Automated DNA Fragments Recognition and Sizing Through AFM Image Processing

Elisa Ficarra[†], Luca Benini[‡], Enrico Macii[†], Giampaolo Zuccheri[◇]
[†] Politecnico di Torino, DAUIN, Corso Duca degli Abruzzi,24 Torino ITALY
[‡]University of Bologna, DEIS, Viale Risorgimento, 2 Bologna ITALY
[◇]University of Bologna, Dep. of Biochemistry and INFM, via Irnerio, 48 Bologna ITALY
{elisa.ficarra@polito.it, enrico.macii@polito.it,
lbenini@deis.unibo.it, giampaolo.zuccheri@unibo.it}

***Abstract*— This paper presents an automated algorithm to determine DNA fragment size from Atomic Force Microscope images and to extract the molecular profiles. The sizing of DNA fragments is a widely used procedure for investigating the physical properties of individual or protein-bound DNA molecules. Several AFM real and computer-generated images were tested for different pixel and fragment sizes and for different background noises. The automated approach allows to minimize processing time with respect to manual and semi-automated DNA sizing. Moreover, the DNA molecule profile recognition can be used to perform further structural analysis. For computer-generated images, the root mean square error incurred by the automated algorithm in the length estimation is 0.6% for 7.8 nm image pixel size and 0.34% for 3.9 nm image pixel size. For AFM real images we obtain a distribution of lengths with a standard deviation of 2.3% of mean and a measured average length very close to the real one, with an error around 0.33%.**

Keywords: DNA sizing, DNA secondary structure transition, molecular profile extraction, image processing, AFM images

I. Introduction

The size of DNA fragments provides essential information for the construction of physical genome maps and genotyping. In particular, by knowing the DNA fragment length and the DNA molecule profile it is possible to investigate the properties of single DNA molecules [6] or of DNA-protein interactions. It is possible, for example, to distinguish between different DNA secondary structures [16] and also to establish if and in which manner a ligand binds to DNA [17].

For DNA sizing, gel electrophoresis methods are very common. Their limitations are the low speed (processing times of 2 hours or more) and large amount of DNA samples required for analysis. An alternative approach resorts to optical microscopy [13], which employs light microscopy to compute the length of fluorescently stained DNA restriction molecules. This method provides good throughput, high resolution and low cost. Accuracy is the same as gel electrophoresis but for smaller DNA fragments optical microscopy is not very effective because resolution is limited to about 600bp [1]. Also, all the methods based on labeling with fluorescent markers achieve worse resolution than optical microscopy; furthermore, they alter the structure of the molecules.

The atomic force microscope scans a solid surface, to which DNA samples in solution have been absorbed, with a sharp probe tip at the end of a flexible cantilever. Voltage measurements from a laser beam deflected off the top of the cantilever to a photodiode are employed to create the images of DNA sample heights. Typically, the atomic force microscope images of DNA have a lateral resolution of 1 to 10nm (while optical microscopy does not go below 200nm). The AFM has a high signal to noise ratio so that it can be applied to biomolecules like nucleic acids. Furthermore, it enables direct visualization of single DNA molecules without contrast-enhancing agents and directly maps the structure of DNA-binding proteins bound to molecules.

Automatic and semi-automatic algorithms for DNA sizing based on AFM have been developed in the recent past. A semi-automatic algorithm is presented in [16]: in an AFM image the edges and several internal points in each fragment must be selected with a manual procedure, a DNA line is traced by interpolating the selected points (anchors) and then this line is skeletonized (thinned) using 8-connectivity. Rivetti [16] reports that among six different methods employed to estimate the contour length of digitalized DNA molecules the best is an order-3 polynomial smoothing of the DNA line over a moving window of 5 points. As a consequence, the computation of the Euclidean distance point by point in the smoothed line achieves an error around 1% with respect to the real contour length. An automatic length estimation algorithm was presented by Spisz et al. [22]. This algorithm uses a set of image processing steps, such as image segmentation, image smoothing with an average filter, image thinning etc. This method processes the images in 1% of the time required by the semi-automated method but is affected by a larger inaccuracy on measured contour length. Another automated algorithm was presented in [19]. Among other steps, the authors proposed a thresholding processing step characterized by a single-fixed threshold. This algorithm too is affected by a larger inaccuracy on measured contour length. Moreover, the effectiveness of this approach has not been proved in presence of high background noise level.

The main contribution of our work is twofold:
(i) design and implementation of a fully automated algorithm which provides same or better accuracy than what is currently obtained using semi-automated approach commonly used by biologists.
(ii) design an automated image processing chain composed by customized processing functions (fragment points recovering, pruning, critical molecules removing, molecule length computation) that provides higher accuracy than previous automated solutions without impacting the execution time (as explained in Section III).

Furthermore, with respect to the semi-automated approach, the fully automated algorithm can process images without any interaction with the operator. This avoids errors introduced from operator bias and increases the amount of information available for further analysis such as DNA intrinsic curvature [12] and dynamic structure analysis, critical for the understanding of several key biological processes (e.g. DNA packaging, transcription, replication, recombination, repair and nucleosome stability and positioning [23]).

II. Length determination algorithm

This section describes an automated algorithm that computes DNA fragment lengths and extracts the molecule profiles from images under varying image conditions, fragment sizes and background

[1]number of base pairs.

contamination. Our algorithm takes as input an AFM image of DNA, computes DNA fragment lengths and profiles through a set of image processing steps and outputs an histogram of sizes and a data file that contains the coordinates and the size of each fragment recognized. This allows to select and isolate every molecule. Our goal is to compute DNA size and molecular profiles with a very high accuracy and low processing time. To extract biological information from AFM images, several processing steps are needed. Each step takes as input the image processed in the previous step in a sequential way. In the following subsections we describe each step composing the algorithm.

A. Filtering

Even if the signal to noise ratio provided by the AFM is higher than those provided by other techniques [3][5][8][21], there is still need of some level of noise filtering. In general, several noise sources can affect the image (i.e. roughness of the support surface, presence of impurity in the sample, different humidity or temperature in the environment, wear of the AFM components). Filtering all these heterogeneous noise sources using an automated approach is impractical due to their variability in nature and intensity. Fortunately, due to their nature, most of these kind of noises are statistically distributed in very small sub-portions of the images. As a consequence, the large part of the image can still be successfully processed. In fact, the noise that most uniformly affects an AFM image is due to a single source that leads to distributed spots[2]. This noise can be classified as impulsive, and thus can be filtered out using a median filter. As a consequence, even if the fragments located in some small portions of the image may not be recognized, this will not affect the effectiveness of the technique, since the very large part of the fragments can be recognized using a fast and automated approach. On the other hand, a deeper analysis of the other noise sources is out of the scope of this paper.

We remove the noise by choosing among a set of filters tuned to the most common AFM image noises. We set as default a 3x3 median filter. With the median filter each output pixel contains the median value in a ordered set of the 3-by-3 neighborhood value. With this kind of filtering, some end points of the fragments or some pixels in the thin fragment areas may be erroneously deleted. Thus, we implemented a procedure that recovers the pixels around the fragments, as explained in Subsection C. We implemented also a Gaussian filter. Both filters work quite well, but the median one defines better the contours of molecules, creates less noise-branches close to the molecules and filters better the noise like striping. We implemented also an adaptive filter. The filter uses a pixel-wise adaptive Wiener method [4] based on statistics estimated from a local neighborhood of each pixel. We used neighborhoods of size 3-by-3 to estimate the local image mean and standard deviation. This filter works better when the noise is constant-power additive noise.

The low-frequency noise in AFM images is a very particular case. It is probably due to uneven support surface. For thoroughness we have inserted a high-pass filter based on Fourier transform analysis between the filters alternative to default.

[2]In fact, the AFM creates the image through a probe (cantilever) that is excited by an electrical oscillator, so that it effectively bounces up and down as it scans over the AFM surface and the sample. Near to high quote zone (like DNA fragments), the probe can oscillate before it reaches the correct level. This is cause of striping or points like 'salt and pepper' in the image. This phenomenon is connected to the frequency of probe oscillation. Decrease them means less precision in following the topography of the sample.

B. Thresholding

This step transforms the original gray-level image in a binary image where pixels labeled '1' represent a possible fragment part. Several techniques have been proposed from researchers in the image processing field to detect a wanted object from the background [18][24][20]. In our recognition framework, we have implemented the Ridler thresholding algorithm [15] that gives us a good trade-off between accuracy and execution time. As detailed in Section III, this method allows to obtain good accuracy without impairing the algorithm's processing speed. We have also compared this method with two others clustering-based thresholding methods, Kittler et al. [9] and Otsu et al. [14]. With Ridler approach we obtained better results than Kittler method and comparable with Otsu one. In fact Ridler method works better than Kittler-thresholding when the image histogram distribution is not bi-modal, as almost always happens in AFM images. See Figure 1. To implement the thresholding processing step, assuming no knowledge about the exact location of fragments, we consider as a first approximation that the minimum gray-level quote pixels are background and the remainder pixels are fragments. At step t we compute the mean background and fragment gray-level, μ_B^t and μ_O^t where segmentation into background and fragments at step t is defined by the threshold value T determined in the previous step as the mean between the mean background and fragment gray-level. The process is straightforward iterative and stops when the threshold value is equal to the previous value (Figure 3.a). Figures 2 and 3 show the output of some of our processing steps. Thus, we refer to this image in almost of following steps.

C. Fragment Points Recovering

After filtering and thresholding, some valid pixels may have been erroneously evaluated as background. In this step, we recover these points. Fragments are identified using the 8-neighbors for connectivity to avoid errors when the fragments are partially or completely aligned on the diagonal connection. Using 8-connectivity we consider neighbors in the vertical, horizontal and diagonal directions for each pixel. In the thresholded image we analyze each neighbor of the fragment pixels. If the neighbor under examination has been evaluated as background after filtering and thresholding, but its value before filtering was over the threshold T, it becomes a valid pixel, hence part of the fragment. We perform this analysis for the neighbors of each fragment pixel as well as the neighbors of each new valid pixel.

D. Thinning

According to the value of neighboring points, this process removes, iteratively and point by point, the pixels of each fragment leaving the skeleton of unit thickness. To extract correct molecular profile, we want to approximate the molecules only with their central axes. Non-maxima suppression after gradient computation has been also tried as an alternative to thinning algorithm. However, since we are interested in finding the backbone of molecules, additional steps will be required after edge computation, thus increasing execution time. Several thinning algorithms can be found in the image processing literature [11][19][10]. Thinning is generally a time-consuming process. Thus, we implemented a simple sequential thinning algorithm so to obtain a good trade-off between accuracy and execution time[3].

For each point in the fragments, we test the match of the point and its neighbors with a set of masks (see Figure 4.a). If the match is

[3]Thinning matches our requirements since we obtain a good overall result on the length computation error. In terms of execution time, thinning is not the critical step, that is pruning and artifact removing, following steps in this section.

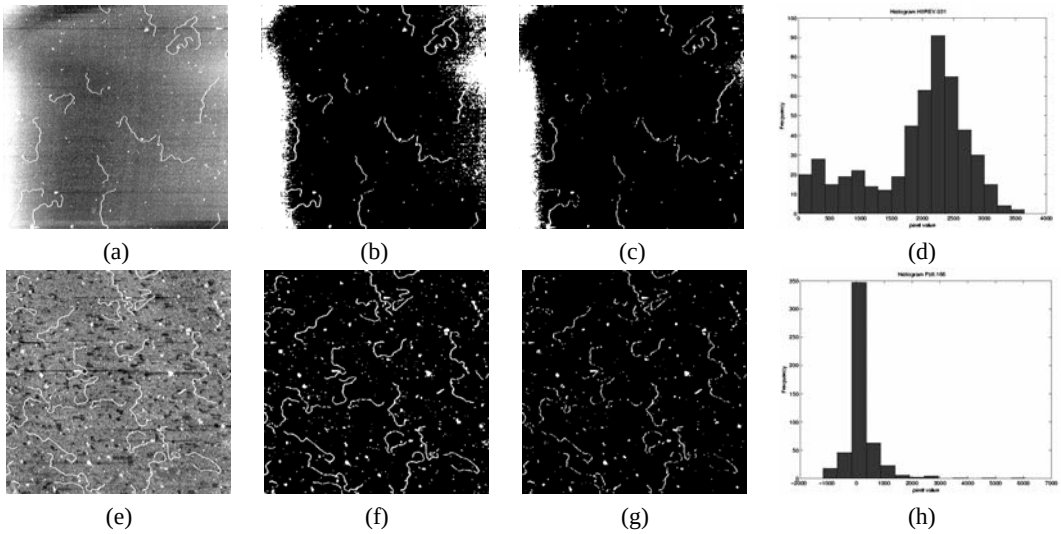


Fig. 1. Comparison between Ridler and Kittler image thresholding methods: (a) original AFM images of EcoRV dimer; (b) post Ridler-thresholding image and (c) post Kittler-thresholding image (note that Ridler method selects more molecules); (d) multi-mode image histogram distribution; (e) original AFM images of PstI dimer; (f) post Ridler-thresholding image and (g) post Kittler-thresholding image (also in this case Ridler method better preserves the connectivity of molecules); (h) single-mode image histogram distribution

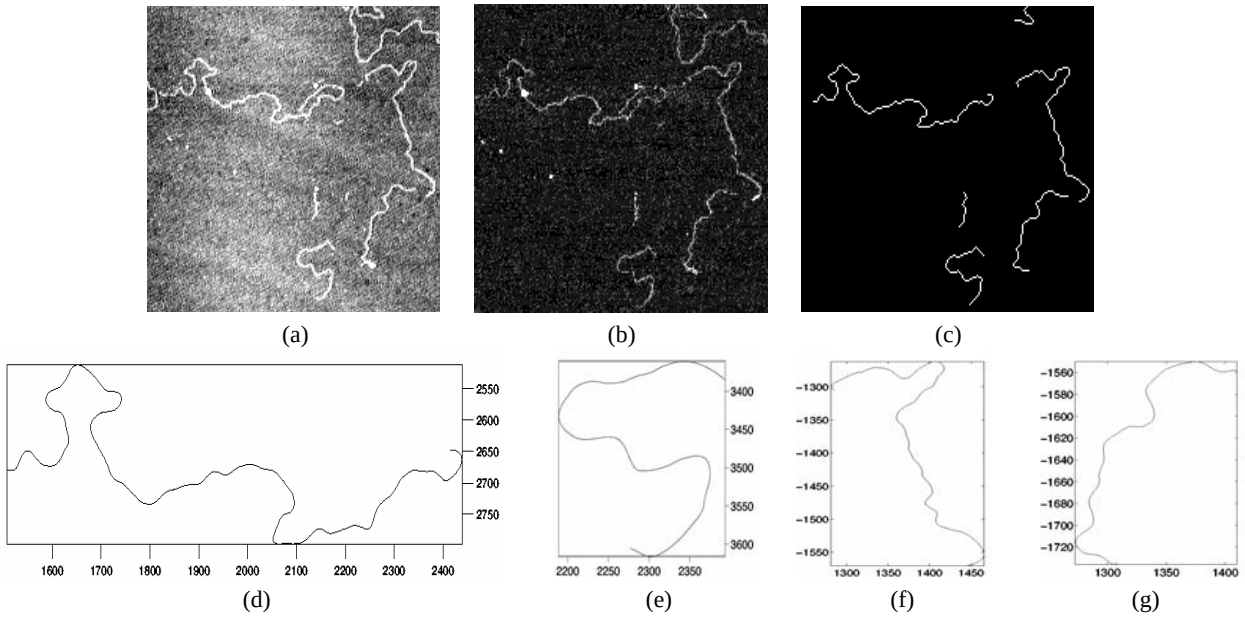


Fig. 2. Example of some image processing step: (a)original image (zoom to center of an AFM image); (b)filtered image (same zoom) using a 3x3 median filter; (c)image at the end of the processing (same zoom); (d),(e),(f) and (g) molecule profiles extracted from the image, the axis represent the coordinates of molecules in the image (nanometers)

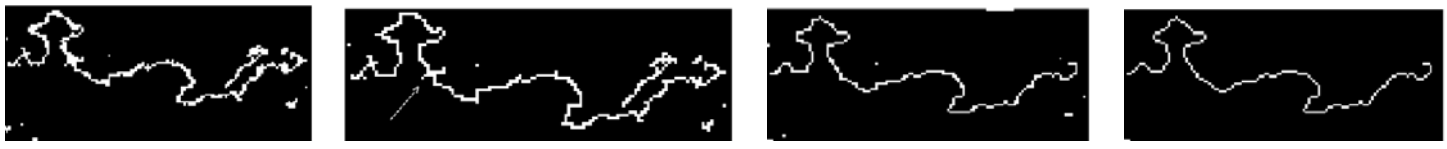


Fig. 3. Example of some image processing steps, zoom on one molecule in an AFM image: (a)molecule after thresholding; (b)molecule after thinning. The arrow in the image (b) shows one of the spurious branches; (c)same molecule after pruning and Critical molecules removing: the isolated fragment on the right in (a) and (b) has been removed in (c); (d)same molecule at the end of the processing

found the point is deleted. This process ends when no more changes in the image are detected. Figure 3.b shows the fragments thinned after this step.

E. Removing objects across the image boundaries

The fragments that are located across the image boundaries must be deleted since it is impossible to determine their extension beyond the image and to completely analyze their features. Note that the presence of molecules located across the image boundaries depends

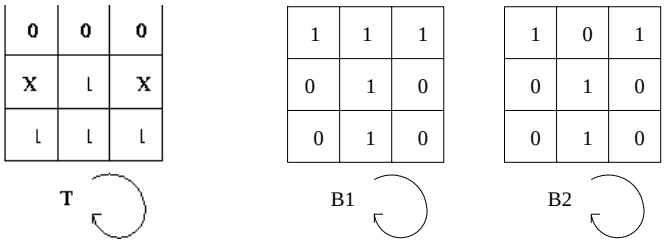


Fig. 4. (a): Example of mask to remove pixels in the *Thinning* step. The central pixel in the T mask is the examined pixel. Using the T mask and its rotations we obtain the possible configurations for deleting the point. The X pixels can have both 1 and 0 values; (b): The B1 8 rotations and the B2 8 rotations are two examples of mask for identifying critical molecules in the *Removing Critical Molecules* step. Specifically B1 and B2 identify molecules that overlap

on the AFM image acquisition process, that is a preprocessing step with respect to our application, and does not depend on the automated or the semi-automated algorithm for DNA sizing we are using. In fact objects across image boundaries are routinely removed by the operator when using manual or semiautomatic procedure.

F. Pruning

We define spurious branches those objects that look as small part of molecules, like branches. They are the result of impurity in the sample dragged by the cantilever tip or due to noise close to fragments that has not been removed in previous steps. Branches differ from main segment because they are much shorter than DNA fragments.

Neighborhood pixels have been analyzed using a set of masks to distinguish among three possible situations, namely spurious branches, critical molecules and corners (we describe critical molecules and corners later in this section). In this step we want to delete branches. For each fragment we recognize the end points matching the molecule pixels with 8 masks.

Thus we recursively delete them from the most externals to the inner ones following the recursion steps and, at the same time, we memorize them in a vector (namely $Endpoints$). The user can fix off-line, as parameter to the algorithm, the depth of the recursion depending on the observation of the image. We defined three levels of recursion: low, medium and high. As an example, for images with an high noise level a high value of recursion will be chosen by the user. This is because the higher the noise, the higher the probability to have longer spurious branches that can be removed with a high number of iterations. The process stops when there are only two end points remaining, even if the depth of recursion is not yet reached[4]. A post processing step recovers the original end points of the fragment without recreating branches (see Figure 3.c). Thus we recover only the pixels that have two main characteristics: (i) the pixels must be end points previously deleted, (ii) the pixels do not make their neighboring pixels end point of molecules. The inputs to this part of code are the two last end points (ind) and the deleted end points vector ($Enpoints$).

G. Removing Critical Molecules

We define the Critical molecules as objects consisting of two or more molecules that overlap, or of a single molecule that overlaps on itself or is closed in circle. In these cases it is not possible to distinguish one molecule from another and to find the end points. Since it is impossible to compute their lengths or to extract each correct profile they should be discarded from further consideration. We studied a set of masks to recognize all these critical cases. Figure 4.b shows some examples. We compare these masks with the neighbors of all points of fragments for identifying the overlap-points and then recursively deleting the molecules. This step has been integrated to the pruning step to further minimize the processing time (see in the Figures 3.b and 3.c the isolated fragment on the right).

[4]A bound on recursion depth is needed to avoid deleting backbone of fragments in case of molecule crossings. In fact, if two molecules overlap a backbone of one molecule can be wrongly interpreted as a spurious branch and thus deleted by pruning. The result of this operation would be a single molecule (obtained by two partially deleted overlapping molecules) that may not correspond to a real fragment, being the composition of two different fragments. We can distinguish real spurious branches from crossing molecules because in general they are due to noise spots close to a molecule, thus they are usually much shorter.

H. Removing Artifacts and corner pixels

Fragments composed by a number of pixels smaller than a user-defined minimum size are considered artifacts and are then deleted. As artifacts we consider noise, like points, or uninteresting sample material, like proteins or different molecules in the same sample. This minimum size is a user-configurable parameter and allows to user to select only the molecules of interest.

The corner pixels are pixels that form an angle of 90 degree with the previous and next pixels. Through neighborhood pixel analysis, also the corner pixels are recognized and deleted since they are due to image pixeling and could introduce some extra error distance during the length computation (see Figure 3.d for artifacts and corner pixels removing).

I. Length Calculation

Several techniques have been proposed from researchers to estimate the length of a digital curve [2][16]. Computationally intensive approaches, such as those based on discrete geometry [1], give a general solution to the problem, but we found that a simpler ad-hoc approach gives us a very good accuracy with a low execution time. Thus, we computed molecule length as the sum of the Euclidean distances between consecutive pixels in the thinned fragment where the pixel coordinates are the horizontal and vertical indices of the pixel in the image. But, except for the edge coordinates that remain unchanged, the other pixel coordinates are calculated as weighed average, using a single weight k, of the previous, current and successive points.

$$X_p = k(x_{p-1} - x_p) + x_p + k(x_{p+1} - x_p) \tag{1}$$

$$L_{p+1,p} = \sqrt{(X_{p+1} - X_p)^2 + (Y_{p+1} - Y_p)^2} \tag{2}$$

Where L is the modified distance between the points with coordinates x and y. The higher the value of k, the higher is the effect of the interpolation between the previous and the next pixel. The smoothing is needed to approximate the real filament shape and length that has been distorted by pixel quantization. In fact, with respect to a simple

Euclidean distance metric, our measurement approach achieves more accurate results since it better adapts to the DNA structure, where the position of a single point (a base) is affected by the position of adjacent points (e.g. bases). Moreover, the skeletonized profile traces the best guessed location of the DNA molecular axis, but due to its pixelized nature it does not approximate well the almost continuous curvature of DNA, and thus the real filament shape and length. To regain this more natural aspect a smoothing operation is necessary prior to length measurement. Finally, the routine has been tested for different weight value and we found values of k that minimize the error on the length measurement (as shown in Section III). The length calculated using the pixel as the unit is scaled to nanometers according to the image pixel size.

J. Molecule Extraction

In this step the molecules are extracted from the image and their pixel coordinates stored in different data files (Figures 2.d, 2.e, 2.f and 2.g). Thus, using a fixed points sliding window each fragment-tract is fitted to a variable-degree polynomial curve that ensures square error smaller than a user-defined threshold. This smoothing is in order to perform further analysis like, for example, DNA intrinsic curvature measurements [12].

III. Experimental Results

The experimental results were obtained by running both real AFM images and computer-generated benchmarks with additive random gaussian noise as background contamination.

A. Implementation

Our algorithm has been implemented using Matlab 6.0 (Release 12) for UNIX platforms. We utilized two MathWorks Image processing Toolbox 2.2.2 functions in the filtering step: the $fft2$ function to compute the two-dimensional fast Fourier transform and the $wiener2$ as implementation of the Wiener adaptive filter.

B. Experimental Set Up

The fragments in the computer generated benchmark images were generated according to Gaussian probability distribution as described in [16] (section 2.3) so as to exhibit similar distribution and similar shape as in the real AFM images of DNA molecules (including tip-broadening effects).

C. Simulated Images

In the computer-generated square images, the tests were organized in two sets depending on pixel size that, due to the image size/pixels number ratio, was set to 7.8 nm (i.e. 2000(nm)/256pixels) and 3.9 nm (i.e.2000(nm)/512pixels).

The first set of results shows the average errors as a function of the fragment length for 512 pixels images. The second set of results shows similar plots for 256 pixels images. For both cases we tested the algorithm with seven different fragment sizes (300, 433, 567, 700, 833, 967, 1100 nm) and for three different additive gaussian noise levels. The additive gaussian noise variance was set to 0.01, 0.02 and 0.06 to obtain these different noise levels. It must be observed that the last noise level is higher than the common noise level in real images. The algorithm has been tested with a thousand fragments for the seven fragment sizes and the three additive noises. Finally, the tests were performed using 21,000 molecules for each of the two 7.8 nm and 3.9 nm pixel size sets. In addition, different values of k were tried in the length calculation procedure in order to find the best value of k (k_{opt}) minimizing the error of the measurements.

As a result, we found two different value for k_{opt} depending on the pixel size. In fact, if the pixel size is larger, k_{opt} should be smaller to compensate for the interpolation inaccuracy (since increasing the value of k enhances the effect of the interpolation, that is more precise for smaller pixel size).

Figures 5.a, 5.b, 5.c show the average absolute errors for each fragment length versus different values of k, in images with additive gaussian noise of mean 0 and variance 0.01, 0.02 and 0.06 respectively. The plots are for pixel size of 7.8 nm.

Figure 5.d show the average absolute errors for each fragment length versus different values of k, in images with additive gaussian noise of mean 0 and variance 0.06 and pixel size of 3.9 nm. As a common behavior for both pixel sizes, it can be observed that for shorter fragments, the best value of k is higher, while the opposite is true for longer fragments. This is because the higher the value of k, the higher is the weight of the interpolation in the length computation formula, hence the error is minimized for most regular fragments, which are normally shorter (shorter fragments frequently have a more regular shape). In fact, when the fragments are longer and have more irregular shape, the effect of the interpolation is to underestimate the length of the molecules[5]. We set k_{opt} as the value that balances the errors in the whole range of fragment lengths, obtaining the minimum average error.

Comparing Figure 5.a with Figures 5.b and 5.c, where the mean of the noise is always 0 but the noise variance is respectively 0.01, 0.02 and 0.06, it can be observed that the error slightly increases as expected when the noise level is higher, but, since the overall behavior does not change for higher noise, the choice of k is independent from the noise level. The same experiments have been repeated for 512 pixel images.

As a result we chose for the 512 pixel images k_{opt}=0.32 and for the 256 pixel images k_{opt}=0.16. The results shown below are obtained for these two values of k_{opt}. Figure 6.a shows thus the results for different fragment length measures for k_{opt} of 0.32 in images with pixel size of 3.9 nm and additive gaussian noise of mean 0 and variance 0.01, 0.02 and 0.06 respectively. For each measure the error is the average among the relative errors computed for the same length in different image fragment number. Positive values of average errors indicate an over-estimation, while negative values indicate an under-estimation of fragment lengths. Figure 6.b reports similar results for the 256 pixel images for k_{opt} of 0.16.

In Figure 7 is shown the root mean square error versus the resolution of image for each length. The error is monotonic function of the pixel size.

Considering all noise and length cases, for 512 pixel images a root mean square error of 0.4% is achieved and for 256 pixel images the root mean square error is 0.62%.

In order to compare our method with the semi-automated [16] one, the latter was used to measure the fragment lengths on the same set of 512 pixel images with noise variance level of 0.06. The root mean square error produced by the semi-automated algorithm on this set of 512 pixel images, is of 1.6%. As a result, our method achieves higher precision (since for 512 pixel images it achieves a root mean square error of 0.4%) with a much shorter processing time (about 60 seconds on a PC equipped with a 650MHz PentiumIII processor compared with about half an hour of the manual procedure, depending

[5]There is an inversely proportional relationship between pixel size and k value. In fact, to keep the same correction on the pixel coordinates (in nm), this should be independent from the pixel size. Now, the correction is proportional to k times the distance among previous and current pixels and among current and next pixels. This distance is larger when pixel size increases. Thus, a lower value of k should be chosen.

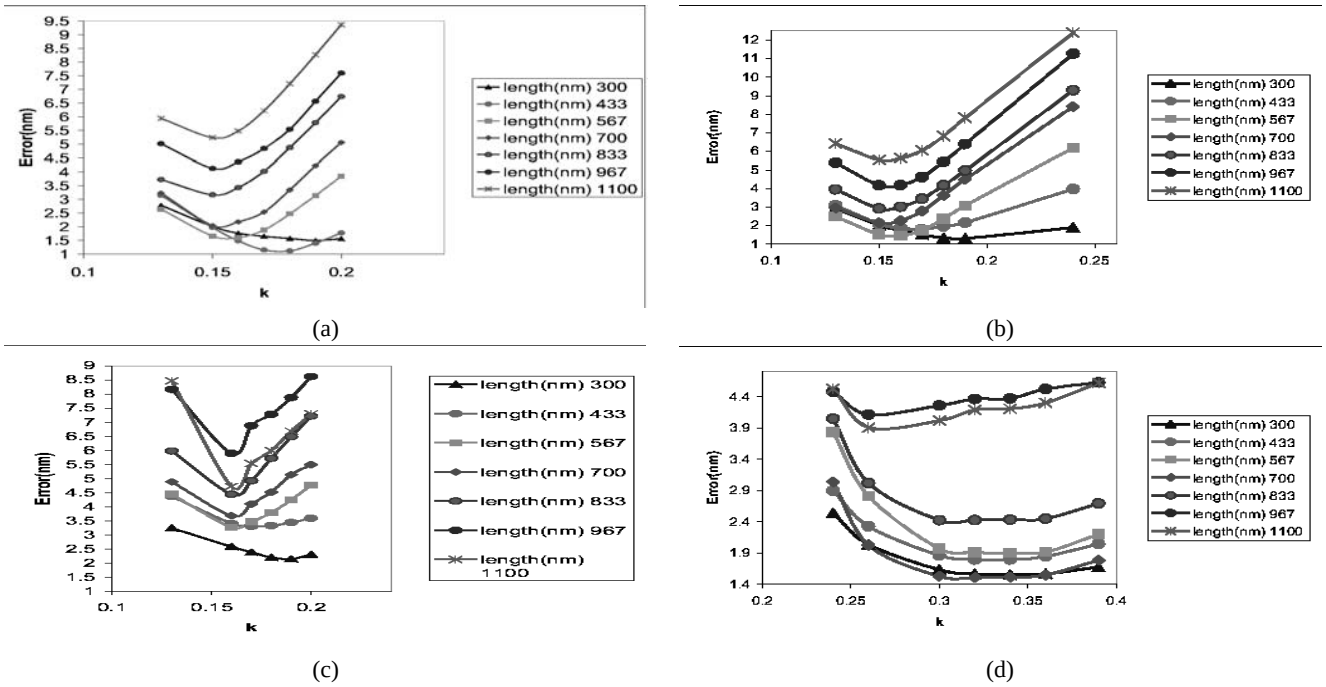


Fig. 5. Absolute length estimation error plots vs. k parameter values; (a), (b) and (c) are absolute length estimation error plots for 0.13, 0.15, 0.16, 0.17, 0.18, 0.19, 0.20 and 0.24 k values in images with additive noise of mean 0 and variance 0.01, 0.02 and 0.06 respectively and pixel size of 7.8 nm. The error is the absolute average among the errors computed for the same length in different images; (d) is always absolute length estimation error plot, but for 0.24, 0.26, 0.30, 0.32, 0.33, 0.36 0.39 k values in images with additive noise of mean 0 and variance 0.06 and pixel size of 3.9 nm. The error is the absolute average among the errors computed for the same length in different images

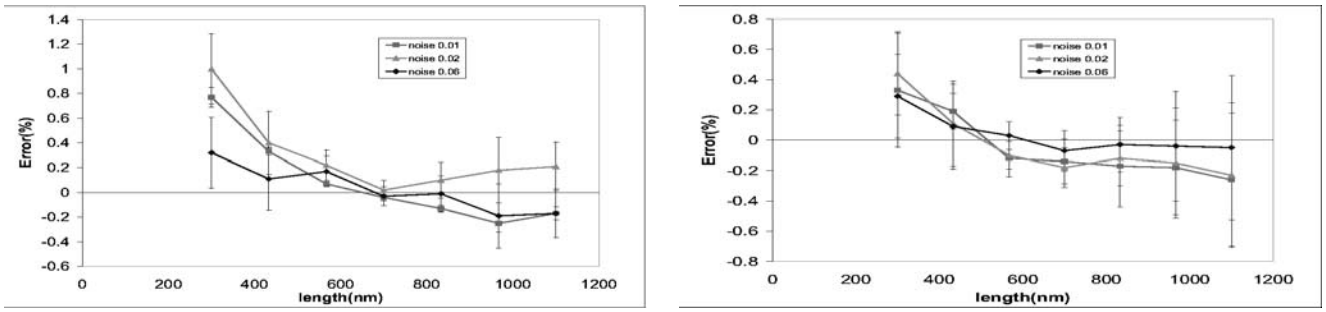


Fig. 6. Error estimation plots in images with additive gaussian noise of mean 0 and variance 0.01, 0.02 and 0.06: (a) error estimation plots for k_{opt} of 0.32 in images with pixel size of 3.9 nm; (b) error estimation plots for k_{opt} of 0.16 in images with pixel size of 7.8 nm. Variances are reported as vertical lines. Note that, since k has been optimized considering all the different fragment lengths and longer molecules have a large probability to have spiky curves, that is where the chosen value of k_{opt} lead to underestimation. On the contrary, smaller fragments are likely to be more flat, and for this reason the value of k_{opt} leads to overestimation. This effect has been corrected finding a linear correlation between the average errors computed and the real fragment lengths, as explained later in this subsection. Note also that if the pixel size is bigger, the variance of the error on the length computation is higher than for 512 pixel images, because of stronger effect of imprecise computation of the number of pixels composing a particular fragment.

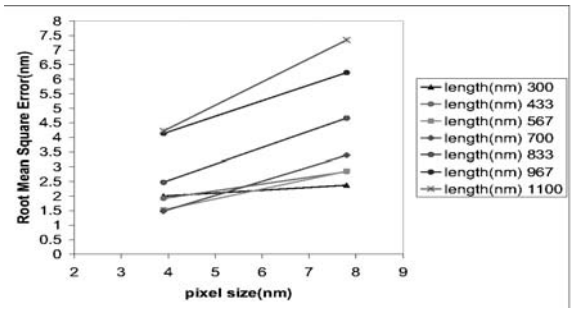


Fig. 7. Root mean square error versus the resolution of image, for each length

to the user skill) [6]. Reducing the processing time is useful to speed up analysis of several images, as this is typically the case in a biological analysis context. In addition, the automated algorithm avoids errors

[6]We obtained this value by asking biologists using semi-automated approaches to give us data about the time spent to complete the process. Timing we use in the paper are thus typical times required by expert users.

introduced from operator bias and it allows to manage a large amount of data to determine useful information such as DNA structure and property analysis.

Examining our results, a linear correlation has been found between the average errors computed and the real fragment lengths with a confidence level of 99%. This allows to compute systematic errors. In order to estimate the confidence level of this correlation, we performed the test *T of Student* under the null hypothesis on the correlation coefficients. As a result, we rejected this hypothesis at significance level less than 1%. Thus the correlation turns out to have a confidence level greater of 99%. As result of the correction, the root mean square error incurred by our algorithm is 0.34% for the 512 pixel images and 0.6% for the 256 pixel images. In this way, we obtained an improvement with respect to the results shown in Figure 6.

In order to compare our method with previous automated algorithms presented in the literature, we performed an additional test with images of molecule length of 90nm. As experiments in previous work [22], the images were at a resolution of 1.953 nm/pixel. At this resolution we had to set the k parameter before to perform the test on the 90 nm molecules. For this reason, the algorithm was tested with fragment sizes of 300, 433, 567 nm and for all the three different additive gaussian noise levels. As result, the k_{opt} was set to 0.48 and the root mean square error incurred by our algorithm was about 0.8%. Finally, we performed the test on a set of 620 molecules of 90 nm of length. We obtained a distribution of lengths with a mean value of 89.3 nm and a standard deviation of 2.2 nm. Thus, the error with respect to the real length is about 0.77% with a standard deviation of 2.4%. As a consequence, our work is found to provide a considerable improvement in accuracy. In fact, in [22] the error on the average length w.r.t. the theoretical DNA fragment size is of 3% and the standard deviation is about 10%.

D. Real AFM Images

Our algorithm was tested with three sets of real images and the results have been compared with those of the semi-automated procedure. All sets were images at a resolution of 3.9 nm/pixel, so k_{opt} was set to 0.32. The molecules in the first two sets of images were 633.4 nm palindromic dimers of DNA obtained by joining two segments in either the head-to-head or the tail-to-tail configuration [26]. The molecules were cut between two different sites (EcoRV and PstI) and dimerized around either site to get two different dimers, EcoRV-EcoRV and PstI-PstI. The third set of images displayed a population of 1098 nm-long DNA molecules containing the 211 bp [7] highly-curved fragment of kinetoplast DNA from $Crithidiafasciculata$.

For each data set, we compared our results with the expected real length value. We performed also a comparison with length distribution obtained with the semi-automated method [16][26]. Generally, a semi-automated procedure can be very effective for selecting molecules of interest because of the skillness of the bio-researcher to distinguish molecules from background noise or artifacts. In particular, the referenced method is widely used in bio-labs [27] because it provides a suitable accuracy. For this reason, we decided to perform also a comparison with it.

The first set of images displayed 170 EcoRV-EcoRV dimer molecules. Figure 8.a shows an example of this set of images. The length distributions of the imaged molecules is shown in Figure 8.b. Although the EcoRV-EcoRV dimer molecules have the same length, the imaged molecules appear to be characterized by somewhat different lengths, as you can see in Figure 8.a. This is due to a partial DNA structural alteration during the deposition process, to residual artifacts on the sample or on the substrate and to the AFM scansion approximations. By observing the length distributions histogram in Figure 8.b it can be noticed that the algorithm selects the molecules of interest removing the artifacts or the critical molecules such as the semi-automated procedure. In fact, comparing our length distributions with the semi-automated measures a similar shape and width in the plots is clearly visible. As a result, we obtained a standard deviation of 2.3%, that is 14.8 nm on a mean value of 631.3 nm. The average length is thus very close to the expected one (633.4 nm) with a 0.33% error.

The second set of images displayed 55 PstI-PstI dimer molecules. Figure 9.a shows an example of these set of images. Just like the previous case, the PstI-PstI dimer molecules appear in the image characterized from some different lengths for the reasons explained above. As for the first set of real images, also for the second one we can notice that the algorithm selects the molecules of interest in good agreement with the semi-automated procedure removing the artifacts or the critical molecules (Figure 9.b). Thus, comparing our length distributions with the semi-automated measures we evidence similar shape and width in the plots. As a result, we obtained a standard deviation of 2.3%, that is 14.6 nm on a mean value of 635.37 nm. The average length is very close to the expected one (633.4 nm) with a 0.3% error.

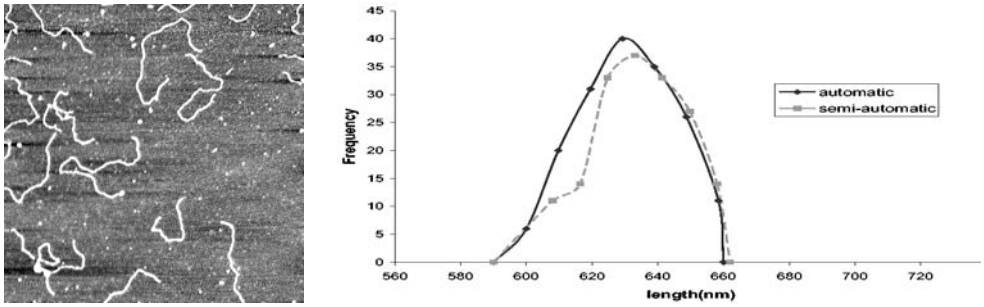
Moreover, comparing these results with previous automated algorithms presented in the literature, our work is found to provide a considerable improvement in accuracy. In fact, for the automated algorithm presented in [22] the error of the average length with respect to the theoretical DNA fragment size for molecules of about same length, is of 1.5% and the precision is of about 5%. For the automated algorithm presented in [19] the error of the average length with respect to the real DNA fragment size is of about 12%. These results are obtained from measures of DNA fragments deposited using air DNA deposition technique, as we did.

The third set of images displayed the 1098 nm-long DNA molecules containing the highly curved fragment of kinetoplast DNA $Crithidiafasciculata$. Figure 10.a shows an example of these set of images. The molecules appear in the image characterized by very irregular profiles due to the unusual very high molecule curvature. This property translates into a harder computation of molecule lengths because surrounding noise shadows DNA shapes proportionally to DNA profile complexity. As for the other sets of real images, also for this one we can notice that the algorithm selects the molecules of interest in agreement with the semi-automated procedure removing the artifacts or the critical molecules (Figure 10.b, solid plot). Comparing our length distributions with the semi-automated measures we evidence similar shape in the plots, with wider spread for the semi-automated one. The average values of the two distributions are very close, but the semi-automated one gives more uncertain information caused by the higher dispersion of the measured lengths. We obtained a standard deviation of 1.9%, that is 20.5 nm on a mean value of 1085 nm. The average length is very close to the expected one (1098 nm) with a 1.18% error. These results are comparable with the two other sets of molecules, considering the unusual structural characteristics of the kinetoplast DNA of $Crithidiafasciculata$ molecule.

IV. Conclusions

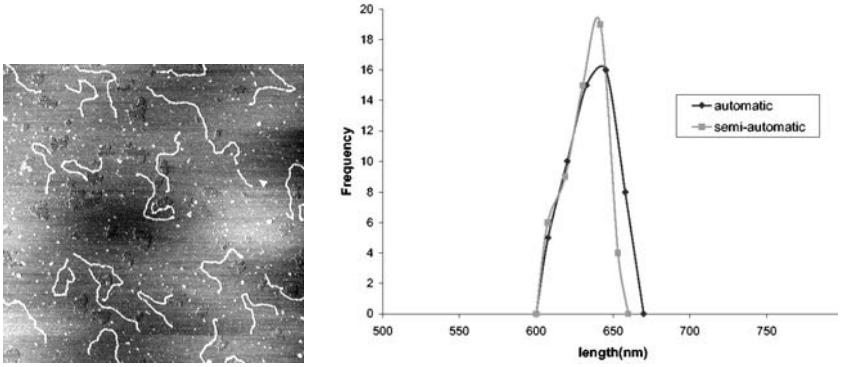
An automated algorithm for determining the DNA molecule length in Atomic Force Microscope images has been presented. Furthermore, the automated approach allows to recognize and extract the molecular profiles with high accuracy, minimizing processing time and increasing the amount of biological and biomedical available information. Our automated algorithm is found to be effective for several DNA

[7] number of DNA base pairs



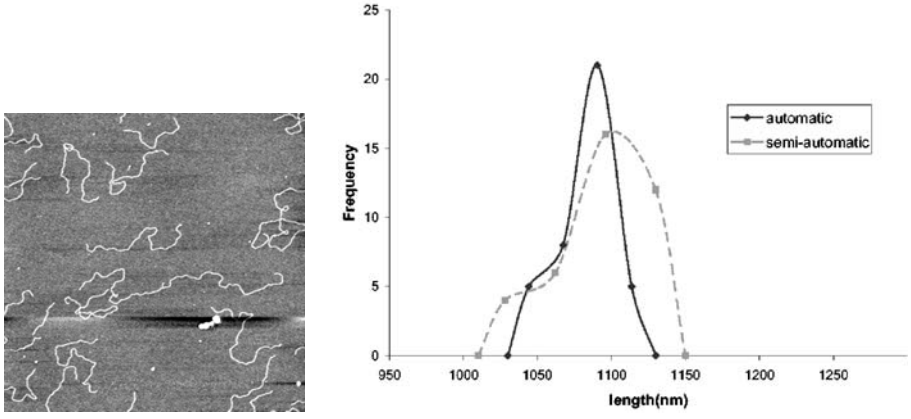
(a) (b)

Fig. 8. (a): EcoRV-EcoRV dimer DNA molecules in an AFM image (particular); (b): Histogram of EcoRV-EcoRV dimer molecules sizing: the solid plot represents our measures, the dashed plot represents the semi-automated measures



(a) (b)

Fig. 9. (a): PstI-PstI dimer DNA molecules in an AFM image; (b): Histogram of PstI-PstI dimer molecules sizing: the solid plot represents our measures, the dashed plot represents the semi-automated measures



(a) (b)

Fig. 10. (a): 1098 nm-long DNA molecules containing the kinetoplast DNA of $Crithidia fasciculata$ in an AFM image; (b): Histogram of $Crithidia fasciculata$ molecules sizing: the solid plot represents our measures, the dashed plot represents the semi-automated measures. The average values of the two distributions are very close, but the semi-automated one presents a higher dispersion of the measured lengths

fragments sizing and molecular profile recognition applications, such as the investigation of transitions in the secondary structure of DNA, interactions between DNA and proteins, static and dynamic structure analysis. The proposed algorithm takes as input AFM images of DNA fragments and elaborates them by a sequence of processing steps. The algorithm has been tested using computer-generated and AFM real images.

Comparing our technique with a semi-automated algorithm, our method achieves higher precision with a much shorter processing time (about 60/120 seconds compared to half an hour). In addition, the automated algorithm avoids errors introduced by operator bias. Moreover, comparing our approach with previous automated methods presented in literature our work is found to obtain a considerable improvement in accuracy. In fact, results on real AFM images show that our algorithm is able to select the molecules of interest by removing the artifacts or the critical molecules as the semi-automated procedure. In particular, we obtained length distributions with a standard deviation of about 2% of mean and an average error of about 0.3% - 1.18%.

As an extension of this work, we are currently designing an automated algorithm for DNA curvature and flexibility analysis. As future work, we plan to implement an adaptive snake algorithm for

increasing the number of selected molecules in very noisy images without impacting accuracy.

V. Acknowledgments

G.Z. wishes to acknowledge support from Progetti Pluriennali Universitị Bologna; FISR D.M. 16/10/20 year 1999 and the ESF Eurocore SONS programme (2003/2006).

References

[1] T. Bulow, R. Klette "Digital Curves in 3D Space and a Linear-Time Length Estimation Algorithm", *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(7): 962-970, July 2003

[2] T. Bulow, B. Yip "Evaluation of curve length measurements", *IEEE International Conference on Pattern Recognition, 2000* 1: 610-613, Sept. 2000

[3] C. Bustamante, C. Rivetti "Visualizing protein-nucleic acid interactions on a large scale with the scanning force microscope", *Annual Review of Biophysics and Biomolecular Structure*, 1996, 25:395-429

[4] K. R. Castleman, "Digital Image Processing", *Prentice-Hall*, Englewood Cliffs, NJ, 1996

[5] Y. Fang, T. S. Spisz, T. Wiltshire, N. D'Costa, I. N. Bankman, "Solid State DNA Sizing by Atomic Force Microscopy", *Anal. Chem.*,70, 2123-2129, 1998

[6] E. Ficarra, D. Masotti, L. Benini, M. Milano, A. Bergia, "Automated DNA Curvature Profile Reconstruction in Atomic Force Microscope Images", *AI*IA Notizie*, 2002, 4, Dec., 64-68

[7] E. Ficarra, L. Benini, B Riccò, G. Zuccheri "Automated DNA Sizing in Atomic Force Microscope Images" *IEEE International Symposium on Biomedical Imaging (ISBI02)* Washington D.C., 453-456, July 2002

[8] H. G. Hansma, J. H. Hoh "Biomolecular imaging with the atomic force microscope", *Annual Review of Biophysics and Biomolecular Structure*, 1994, 23:115-139

[9] J. Kittler, J. Illingworth, C. Y. Suen "Minimum error thresholding", *Pattern Recognition* 19: 41-47, 1986

[10] L. Lam, S. W. Lee, C. Y. Suen "Thinning methodologies-a comprehensive survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14(9): 869-885, 1992

[11] C. M. Ma, M. Sonka, "A fully parallel 3D thinning algorithm and its applications", *Computer Vision and Image Understanding*, 64:420-433, 1996

[12] D. Masotti, E. Ficarra, E. Macii, L. Benini, "Techniques for Enhancing Computation of DNA Curvature Molecules", *IEEE Fourth Symposium on Bioinformatics and Bioengineering (BIBE2004* Taichung, Taiwan, May 2004

[13] X. Meng, K. Benson, K. Chado "Optical mapping of lambda bacteriophage clones using restriction endonucleases", *Nature Genetics*, 9, 432-438

[14] N. Otsu "A threshold selection method from gray level histograms", *IEEE Trans. on Systems, Man, and Cybernetics*, SMC-9: 62-62, 1979

[15] T. W. Ridler, S. Calvard, "Picture thresholding using an iterative selection method", *IEEE Trans. on Systems, Man, and Cybernetics*, 8(8): 630-632, August 1978

[16] C. Rivetti, S. Codeluppi, "Accurate length determination of DNA molecules visualized by atomic force microscopy: evidence for a partial B- to A-form transition on mica", *Ultramicroscopy*, 2001, 87, 55-66

[17] C. Rivetti, M. Guthold, C. Bustamante "Wrapping of DNA around the E.coli RNA polymerase open promoter complex", *EMBO J.*, 1999, Vol.18 (16), 4464-4475

[18] P. K. Sahoo, S. Soltani, A. K. C. Wong, Y .C. Chen, "Survey of thresholding techniques", *Computer Vision, Graphics and Image Processing*, 41(2):233-260, 1988

[19] A. Sanchez-Sevilla, J. Thimonier, M. Marilley, J. Rocca-Serra, J. Barbet "Accuracy of AFM measurements of the contour length of DNA fragments adsorbed on mica in air and in aqueous buffer", *Ultramicroscopy*, 2002, 92, 151-158

[20] M. Sezgin, B. Sankur "A Survey over Image Thresholding Techniques and Quantitative Performance Evaluation", *Journal of Electronic Imaging*, Vol. 13(1), 146-165, Jan. 2004

[21] Z. Shao, J. Mou, D. M. Czajkowski et al. "Biological atomic force microscopy: what is achieved and what is needed", *Adv. Phys.*, 1996, 45:1-86

[22] T. S. Spisz, Y. Fang, R. H. Reeves, C. K. Seymour, I. N. Bankman, J. H. Hoh, "Automated sizing of DNA fragments in atomic force microscope images", *Med.Biol.Eng.Comput.*, 1998, 36, 667-672

[23] Travers, A. A "DNA-Protein Interactions" (Chapman and Hall, London), 1993

[24] J. F. Wang, P. J. Howarth, "Edge following as graph searching and Hough transform algorithms for lineament detection", *Proceeding of IGARSS 89*, Vancouver, Canada, 93-96, IEEE, New York, 1989

[25] T. Y. Zhang, C. Y. Suen, "A fast parallel algorithm for thinning digital patterns", *Communications of ACM*, Vol 27, Num 3, March 1984

[26] G. Zuccheri, B. Samorì "Scanning Force Microscopy Studies on the Structure and Dynamics of Single DNA molecules" *Methods in cell biology*, Vol 68 (Chapter 17) :357-95, 2002

[27] G. Zuccheri, A. Scipioni, V. Cavaliere, G. Gargiuolo, P. De Santis, B. Samorì "Mapping the intrinsic curvature and flexibility along the DNA chain" *Proc Natl Acad Sci U S A*, Vol 98(6) :3074-9, March 2001

Elisa Ficarra Elisa Ficarra received the Laurea degree in Electrical Engineering with specialization in Bioengineering from the University of Bologna, Italy, in 2001. She is PhD. student in control and computer engineering at the Politecnico di Torino under the supervision of Prof. Enrico Macii. From 2001 she is also research assistant at the Department of Electrical Engineering and Computer Science of the University of Bologna under the supervision of Prof. Luca Benini. In 2002 she was visiting at the Stanford University in the Computer Systems Laboratory (CSL). Now she is intern at the EPFL of Lausanne, Switzerland, Facult de Informatique et Comunications. Elisa Ficarra's research interests include computer vision, biomedical and molecular imaging, gene expression analysis, gene clustering and networks.

Luca Benini Luca Benini is an Associate Professor at the Department of Electrical Engineering and Computer Science (DEIS) of the Universty of Bologna. He received a Ph.D. degree in electrical engineering from Stanford University in 1997. Dr. Benini's research interests are in all aspects of computer-aided design of digital circuits, with special emphasis on low-power applications, and in the design of portable systems. On these topics he has published more than 250 papers in international journals and conferences and three books. He has been program chair and vice-chair of Design Automation and Test in Europe Conference. He is a member of the technical program committee and organizing committee of several technical conferences, including the Design Automation Conference, International Symposium on Low Power Design, the Symposiom on Hardware-Software Codesign.

Enrico Macii Enrico Macii holds a Dr. Eng. degree in Electrical Engineering from Politecnico di Torino, Italy, a Dr. Sc. degree in Computer Science from Universita‘ di Torino, and a Ph. D. degree in Computer Engineering from Politecnico di Torino. From 1991 to 1994 he was an Adjunct Faculty at the University of Colorado at Boulder. Currently, he is a Full Professor of Computer Engineering at Politecnico di Torino. His research interests include several aspects of the computer-aided design of integrated circuits and systems. He has authored over 250 journal and conference articles in the areas above, including a paper that received the Best Paper Award at the 1996 IEEE EuroDAC conference. Enrico Macii is an Associate Editor of the IEEE Transactions on CAD (since 1997) and an Associate Editor of the ACM Transactions on Design Automation (since 2000). He was the Technical Program Co-Chair of the IEEE Alessandro Volta Memorial Workshop on Low Power Design in 1999, the Technical Program Co-Chair and the General Chair of the ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED) in 2000 and 2001, respectively, the General Chair and the Technical Program Chair of the PATMOS workshop in 2003 and 2004, respectively.

Giampaolo Zuccheri Giampaolo Zuccheri is full-time Staff Researcher at the Department of Biochemistry of the University of Bologna (since 2002). He holds a degree in Industrial Chemistry (Univ. of Bologna) and a Ph.D. in Chemistry (Univ. of Calabria). He has worked at the Lawrence Berkeley National Labs (Berkeley, California) and at the University of Oregon. He is currently working with Bruno Samor and teaching a class in nanobiotechnology for the degree in biotechnology of the University of Bologna. His interests focus on the chemistry and biophysics of nucleic acids and proteins and on their nanobiotechnological applications. In 1994, Dr. Zuccheri was one of the recipients of the annual prize of the Italian Federation of the Chemical Industry (Federchimica) and in 1998 he was awarded the Borsellino prize of the Italian Society for Pure and Applied Biophysics (SIBPA). He is currently a member of the National Institute for the Physics of Matter (INFM), of the Italian Chemical Society (SCI), of the National Consortium of Materials Science and Technology (INSTM).