Automatic Intrinsic DNA Curvature Computation from AFM Images

Original
Automatic Intrinsic DNA Curvature Computation from AFM Images / Ficarra, Elisa; Masotti, D; Benini, L; Macii, Enrico; Zuccheri, G; Samori, B.. - In: IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING. - ISSN 0018-9294. - 52(12):(2005), pp. 2074-2086. [10.1109/TBME.2005.857666]

Publisher:
IEEE

Published
DOI:10.1109/TBME.2005.857666

Automatic Intrinsic DNA Curvature Computation from AFM Images

Elisa Ficarra†, Daniele Masotti†, Enrico Macii†, Luca Benini‡, Giampaolo Zuccheri⋄, Bruno Samorı⋄

† 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, daniele.masotti@polito.it, enrico.macii@polito.it lbenini@deis.unibo.it, giampaolo.zuccheri@unibo.it, bruno.samori@unibo.it}

Abstract

Critical information on several biological processes such as DNA-protein interactions and DNA transcription can be derived from analysis of DNA curvature. Under thermal perturbation, the curvature is composed of static and dynamic contributions, thus can be described as the sum of *intrinsic curvature* and a fluctuation contribution. Without considering thermal agitations, the DNA curvature is reducible to the intrinsic component, which is a function of the DNA nucleotide sequence only.

In this work we present an automated algorithm to determine the DNA intrinsic curvature profiles and the molecular spatial orientations in Atomic Force Microscope images. The algorithm allows to reconstruct the intrinsic curvature profile by filtering the thermal contribution. It detects fragment orientation on AFM images without labels with a percentage of correct molecular-orientation detection of 96.79% in computer-generated benchmarks, for molecules with a high curvature peak. The automated algorithm reconstructs the intrinsic curvature profile of DNA molecules with a mean square error of $3.8122 \cdot 10^{-4}$ rads over a profile with a central peak value of 0.196 rads, and $6.1 \cdot 10^{-3}$ rads over a curvature profile with two symmetric peaks of about 0.08 rads. Moreover, it correctly detects the location of the peaks in the molecules with a deviation of about 1% of molecule length.

Keywords: DNA curvature, DNA secondary structure transition, non-linear optimization, AFM images

I. Introduction

A precise quantitative knowledge of the local curvature and flexibility is crucial to understand the molecular biology of DNA-protein interactions. In fact, it has been widely recognized that the local sequence-dependent curvature and dynamics of the DNA chain segments play a crucial and active role in DNA packaging, transcription, replication, recombination, and repair processes and in nucleosome stability and positioning [1][2][3]. The experimental study of intrinsic DNA curvature and flexibility has been performed using many different approaches, such as x-ray crystallography [4][5], electron microscopy [6][7], scanning force microscopy (SFM) [8][9], gel retardation, and circularization kinetics [10][11] have been used thus far to study the effects of intrinsic DNA curvature and flexibility. Most of the methods have dealt with large populations of molecules, for which the knowledge of the curvature fluctuations of the individual molecules is blurred in the average population behavior and extracted indirectly from the comparison of many data sets (for example, in approaches using gel-electrophoresis or cyclization kinetics [12][13]).

Recently, DNA curvature has been evaluated from data coming from single DNA molecules, so that the curvature deviation of a molecule from the population average can be estimated [14]. This single-molecule approach is possible with high-resolution microscopy techniques, such as the Atomic Force Microscope (AFM). From AFM micrographs, it is also possible to estimate the local curvature of a section of a molecule, to compare it with the average molecule curvature or the population-averaged curvature of that section [15][16]. Thus, it is possible to obtain a direct estimate of the curvature vector along the DNA chain, once the molecule has been adsorbed on a flat substrate using a semi-automated method. The experiments can be performed in a variety of environments and for DNA molecules of virtually any sequence; furthermore, this type of analysis can be adapted to work on DNA molecules involved in DNA-protein interactions or other biologically relevant functions [17][18].

Curvature analysis on AFM images of DNA molecules is usually performed by skilled operators in a semi-automated fashion. This is understandably a very time-consuming, error-prone and labor-intensive process. In this work we present a fully automated method that performs the curvature analysis with the same level of accuracy. To the best of our knowledge, this is the first automated algorithm implemented to this purpose. In particular, we introduce an algorithm for automatically computing the curvature of DNA molecules starting from Atomic Force Microscopy (AFM) images.

At the microscopic level of resolution, DNA molecules can be idealized as one-dimensional curved lines because of their highly asymmetric form factor. The curvature along the DNA chain is a vector function of nucleotide sequence and is defined as the angular deviation of the central backbone between two consecutive base pairs. Thus, it is possible to describe this term with the function $C(n)$, where n is the base number of the nucleotide sequence. The curvature is composed by static and dynamic contributions. The first one is called *intrinsic curvature* and is a function of the DNA nucleotide sequence only. The second one is caused by thermal energy. The susceptibility to thermal deformation is measured by the *flexibility* of the molecules; this is also a function of the nucleotide sequence [15]. Under thermal perturbation, the observed curvature can be described as the sum of intrinsic curvature and fluctuation contribution $f(n)$.

Thermal perturbation can be seen as a strong noise source, which commonly deforms the molecules with respect to the profile extracted from their intrinsic curvature and it prevents from recognizing particular patterns or defining effective similarity measures among two series of values. For this reason, curvature computation for DNA samples is a non-trivial task.

In the proposed algorithm, curvature analysis is performed in two steps. First, the AFM image is processed to identify DNA molecules and to extract their profile. Then, the extracted molecular profiles are analyzed for curvature computation. The main issue in the second step is to reconstruct the DNA intrinsic curvature profiles through the determination of the correct spatial orientation of each molecule on the AFM substrate following the DNA adsorption process. This second step is the core of our algorithm and it can be handled as a combinatorial optimization problem.

The correct identification of molecule orientation allows to compute curvature in the same points along the chain over a population of same molecules. Unfortunately, in an AFM image it is impossible to distinguish the beginning from the end of a DNA molecule because they lack distinctive features. In the past, the different termini of a sequence has been identified by introducing a tag on one end of the molecules [19]. Nevertheless, it is more conceptually sound to avoid the use of end labeling, since the presence of a bulky or otherwise structurally distinctive object can change the mobility of the entire molecule or of its termini, through the enhancement of self-avoiding effects or through differential molecule-surface interactions that might confer a lower or higher mobility to the tagged section of the molecule. In the past, it has been shown that protein tagging might be used safely [8], but it seems more general and safe (and experimentally simpler) to develop analysis methods that do not employ tags.

The detection of the molecular orientation in the image without any tag is unavoidably a non-deterministic process, and the quality of the decision process should be assessed in a statistical sense. With respect to this issue we propose a fast orientation-finding algorithm that modifies one molecular orientation at a time (fragment flipping) with linear-time heuristic transitions.

We first compare our greedy heuristic will well-known general-purpose heuristic solvers, and we show that it achieves better results with lower computational effort. Finally, we compare the intrinsic curvature profiles obtained with our algorithm against those achieved by a theoretical approach referenced by many other papers in the field [20][21] and taken for comparison[16]. Results demonstrate the high accuracy of our approach.

The paper is organized as follow: In Section II, we survey our molecular profiles extraction algorithm from the images. In Section III we describe in details the fragment flipping and intrinsic curvature profile reconstruction algorithm. Section IV is dedicated to the experimental validation of our approach; finally, conclusions are drawn in Section V.

II. Length determination and molecular profiles extraction Algorithm

A. Automated AFM image processing and length computation algorithm

Atomic Force Microscopy is characterized by high resolution and high signal-to-noise ratio, so it can be applied to nucleic acids. Furthermore, it allows direct visualization of single DNA molecule without contrast-enhancing agents and thus without altering the structure of the molecules. For these reasons it is very effective and useful to study the features of the DNA molecules. The AFM images are 2.5D topographs coded as gray-level images.

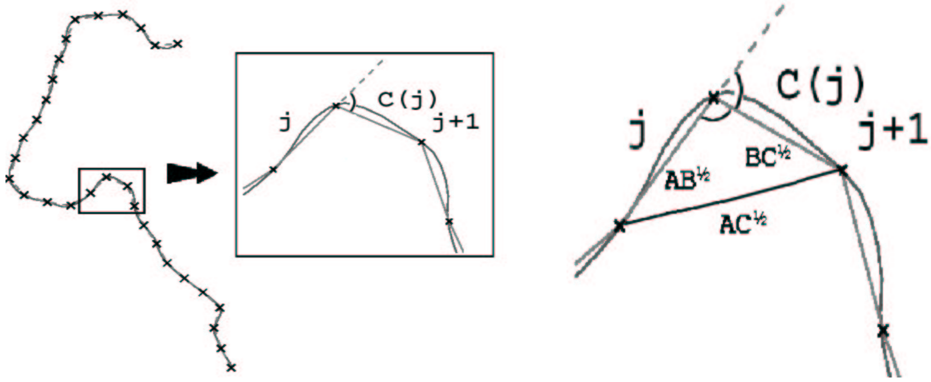
We start by describing the first step in the intrinsic curvature profile reconstruction process. The algorithm computes DNA fragment lengths and extracts molecule profiles from images. In particular, the computation of the lengths is useful to verify that DNA imaged molecules have not undergone a structural transition after deposition on the AFM substrate. In fact, DNA dehydration can drive the transition B- to A-DNA in the secondary structure [22][23]. Researchers conclude that DNA retains generally its B structures on the AFM substrate. But, depending on the characteristics of the dehydration process, such as time for drying, residual humidity of the sample, adhesion of DNA on the surface, it was found that DNA molecules imaged after dehydration can be somewhat shorter than expected for B-DNA. This phenomenon can be caused by a partial B- to A-transition which is more probable for shorter molecules [18] [24]. Transition B- to A-DNA in the secondary structure can be detected on the basis of the measurement of the contour length. In fact the A-DNA form is shorter than B-DNA of about

30%.

Our algorithm, in this step, takes as input an AFM image of DNA, computes DNA fragment lengths and profiles through a set of image processing steps built in a sequential way. This allows selection and isolation of every molecule in order to perform curvature computations. In this section, we survey our length determination and molecular profile extraction algorithm. For more details see [25] [26]. The individual steps are:

- *Filtering*: Even if AFM has a signal to noise ratio higher than those provided by other techniques [17][27] there is still need of some level of noise filtering. The noise in the images is strongly correlated to the experimental conditions and these are not the same ones, not even for the same DNA sample. Filtering all these heterogeneous noise sources using an automated approach is impractical due to their variability in nature and intensity. However, the noise that most uniformly affects an AFM image is due to sources that lead to distributed spots (i.e. localized noise as impurity in the sample and probe-induced noise, as probe wear). This noise can be classified as impulsive, and thus can be filtered out using a median filter. Other filters like gaussian, wiener adaptive, frequency domain analysis based filters were implemented and they can be chosen according to particular cases of AFM image noises. We set the median filter as default.
- *Thresholding*: This step transforms the original gray-level image to a binary image, where pixels labeled '1' represent a possible fragment part. Our algorithm implements the thresholding procedure described by Ridler [28].
- *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 unitary thickness.
- *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.
- *Pruning and Critical Molecules Removing*: It is possible to find some spurious branch close to molecules. These branches could result from impurity or noise close to fragments and not removed in previous steps. Spurious branches are much shorter than fragment length. In this step, we recognize these branches for each fragment and recursively delete them [26].

  We delete also objects consisting of two or more molecules that overlap or as a single molecule self-overlapped or molecules closed in circle, namely *Critical cases*. In these cases, the molecules have to be discarded because it is not possible to distinguish the correct molecular profiles or to find the real end points [26]. Both pruning and critical molecule removing step do not affect the molecule length and the molecular profile.
- *Removing Artifacts and Length Calculation*: We delete fragments with number of pixels smaller than a user-defined minimum size that signify residual noise or not interesting sample material, namely *artifacts*. This minimum size is a user-configurable parameter and allows to user to select only the molecules of interest. To compute the molecule lengths, we define the pixel coordinates such as the horizontal and vertical indices of each pixel in the image. Except for the fragment edge coordinates that remain unchanged, the other pixel coordinates are recalculated as weighted average, using a single weight k, of the previous, current and following points as



(a) (b)

Fig. 1. (a): Example of curvature samples computation; in particular, example of curvature computation of the j-th sample along a DNA chain. We represented two lines: the spline fitting of smoothed experimental molecular profile (curved plot) and the segmental chain (segmented plot) where the number v of points was chosen small enough to easily distinguish the two plots in the figure; (b): Curvature value *C(j,m)*, the curvature is computed from jth and (j+1)th m bp segment coordinates by applying the theorem of Carnot to calculate its supplementary angle *alpha*

shown in Equation 1. The molecule length is then calculated as the sum of the Euclidean distances between consecutive pixels (see Equation 2).

$$X_i = k(x_{i-1} - x_i) + x_i + k(x_{i+1} - x_i) \tag{1}$$

$$L_{i+1,i} = \sqrt{(X_{i+1} - X_i)^2 + (Y_{i+1} - Y_i)^2} \tag{2}$$

where L is the modified distance between the points with original coordinates x and y.

This approach allows us to achieve an improvement in accuracy w.r.t. other length measurement methods [22] and w.r.t. previous automated DNA length computation presented in the literature [29] [30], as discussed in our work [26]. 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 of local curvature computation. Finally, we found values of k that minimize the error on the length measurement. Such values only depend on image resolution, namely the pixel size. This makes sense since the higher the value of k, the higher is the effect of the interpolation between the previous and the next pixel, k should be smaller if the pixel size is larger, to compensate for the interpolation inaccuracy (the interpolation is more precise for smaller pixel size). The length calculated using the pixel as the base unit is scaled to nanometers according to the image pixel size.

B. Molecule Extraction and Curvature sample values Computation

The molecules are extracted from the image and their pixel coordinates stored in different data files. To obtain curvature samples, the molecule profiles have been smoothed along the DNA chain as a fitting of splines from experimental points to guarantee a root mean square error smaller than a user-defined threshold. From experiments we set to 0.3nm the threshold value.

Due to AFM resolution limits, DNA molecules coming from the images are fitted by segmental chains. Thus, we found out v points along each smoothed profile, in order to obtain $(v + 1)$ segments standardized based on their length. Each segment is m bp (DNA base-pair), depending on the micrograph resolution.

The curvature samples were computed as the angles between nearest-neighbor chain segments. Thus, the curvature $C(j, m)$ in Figure 1.a represents the angular deviation between the local profile of the jth and $(j + 1)$th m bp segment. Equation 3 is used for curvature computation, in agreement with [16].

$$C(j, m) = sign \star (\pi - acos(cos(alpha))) \tag{3}$$

where *acos(cos(alpha))* is the arccosine of *cos(alpha)* and *sign* represents the sign of the curvature in accordance with the following convention: the value is positive for counter-clockwise (CCW) angles.

If x and y are the coordinate vectors of the points along the molecules and $(i - 1)$, i, $(i + 1)$ the three molecule points identifying the jth and $(j + 1)$th m bp segments, we can define *cos(alpha)* in accordance with the theorem of Carnot [32] as

$$cos(alpha) = \frac{AB + BC - AC}{2 \star \sqrt{AB \star BC}} \tag{4}$$

where AB, BC, AC are

$$AB = (x_i - x_{i-1})^2 + (y_i - y_{i-1})^2$$

$$BC = (x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2$$

$$AC = (x_{i+1} - x_{i-1})^2 + (y_{i+1} - y_{i-1})^2$$

which represent the square of the jth and $(j + 1)$th m bp segments and the square of the segment opposite to the $alpha$ angle (see Figure 1.b), respectively. From $(v + 1)$ segments we obtain v curvature samples for each DNA fragment.

III. Fragment flipping Algorithm

The curvature is composed of static and dynamic contributions. The first one, called *intrinsic curvature*, represents the interactions between consecutive base pairs only. The second one, called *flexibility*, represents the thermal perturbation due to the interactions of molecules with the environment. Flexibility is also a function of the nucleotide sequence.

Given the base number n of nucleotides sequence, the observed curvature can be described as $C(n) = C_0(n) + f(n)$, where $C_0(n)$ indicates the intrinsic curvature and $f(n)$ denotes the fluctuation terms.

The thermal noise imposes variations on the molecule structure with a contribution characterized by a null mean value since the thermal fluctuations are considered to follow the first-order elasticity because of the relatively high rigidity of DNA [31]. Thus, the structural dynamic perturbations have an average value $< f(n) >= 0$. This leads the estimation of the intrinsic curvature profile considering the curvature profiles of a significant population of molecules with the same nucleotides sequence. Averaging on the population along the chains at position n we obtain the intrinsic curvature value $C_0(n) =< C(n) >=< C_0(n) > + < f(n) >$, while standard deviation value of $C(n)$ is used to characterize $f(n)$.

From an AFM image with m equal molecules, v curvature values are sampled at regular intervals along each chain. Since all these curvature vectors have the same dimension v, we can define a curvature matrix $C(m \times v)$ where element at row i and column j represents the j^{th} curvature value of the i^{th} molecule.

In the DNA deposition process, the molecules can assume different orientations on the substrate. Since the molecular profiles are always extracted by scanning the image left-to-right and up-to-down, we need to know the orientations of the molecules in the image in order to identify corresponding points along the molecules. Thus, we have to represent molecules in the matrix with the same orientation to evaluate the curvature average on corresponding points of the molecule.

The molecules on the image can be characterized by four orientations[1]. Because of different orientations, curvature values in the matrix are the result of curvature sampling on the molecules in one of the four directions: left to right, right to left, up to down and down to up. It is possible to transform each of these orientations to any other by changing the series of curvature values, i.e. inverting the sign or reversing the order of curvature samples.

When we consider real data, single-pair curvature vector comparison is a difficult task. The thermal perturbation can be seen as a strong noise source, that commonly deforms the molecules with respect to the profile extracted from their intrinsic curvature and it prevents from recognizing particular pattern or to define effective similarity measures among two series of values. The optimal configuration, when all the molecules share the same orientation, is the one in which we can observe the minimal value of curvature variance for each point, i.e. the minimal column variance in matrix C.

Because of noise effects given by thermal perturbation, the optimal state column of variance values will not be null. Variance in a point is expected to be the square of the flexibility f in that point. It is important to stress that for the optimal state column is minimum with respect

[1] The planarity of DNA molecules with relevant peaks of curvature leads to the deposition of the molecules in two ways. Moreover, for each one of these two orientations, the molecules can assume two other orientations on the substrate due to the asymmetry of the DNA

to other possible orientations. Following this consideration, the metric chosen to define the state optimality is the mean value of columns variances:

$$\min_{S^j} M = \frac{1}{v}\sum_{j=0}^{v-1}\sigma_j^2 \tag{5}$$

where S^j is the space of the all possible row orientations, σ_j^2 is the j column variance. According to the definition of variance, the objective function M can be written as

$$\frac{1}{v}\sum_{j=0}^{v-1}\frac{1}{m-1}(\sum_{i=0}^{m-1}c_{ij}^2 - m(\sum_{i=0}^{m-1}c_{ij})^2) \tag{6}$$

where c_{ij} are the C matrix elements. Equation 6 can be expanded as

$$\frac{1}{v(m-1)}\sum_{j=0}^{v-1}\sum_{i=0}^{m-1}c_{ij}^2 - \frac{m}{v(m-1)}\sum_{j=0}^{v-1}(\sum_{i=0}^{m-1}c_{ij})^2) \tag{7}$$

The different orientations of the single molecule r, with $r = 1 \cdots m$, can be expressed by using two degrees of freedom, (i.e., 0-1 decision variables) which correspond to inverting the sign of elements of row r or reversing the element order.

Let define four different Boolean decision variables x_{ra}, x_{rb}, x_{rc}, x_{rd} for each row r, each one represents a possible orientation of the respective molecule, i.e. :

$$x_{ra} = \begin{cases} 1 & if\ row\ r\ is\ unchanged \\ 0 & otherwise \end{cases}$$

$$x_{rb} = \begin{cases} 1 & if\ row\ r\ has\ changed\ the\ order\ only \\ 0 & otherwise \end{cases}$$

$$x_{rc} = \begin{cases} 1 & if\ row\ r\ has\ changed\ the\ signs\ only \\ 0 & otherwise \end{cases}$$

$$x_{rd} = \begin{cases} 1 & if\ row\ r\ has\ changed\ both\ sign\ and\ order \\ 0 & otherwise \end{cases}$$

Figure 2 shows an example of transformations of the curvature matrix and the corresponding values of the Boolean decision variables. The first matrix represents a set of molecules (4 molecules) in all the possible orientations, the second matrix represents the same set with all the molecules that have been led to the same orientation.

We use four Boolean decision variables instead of the two that are strictly necessary for the two degrees of freedom, in order to relate

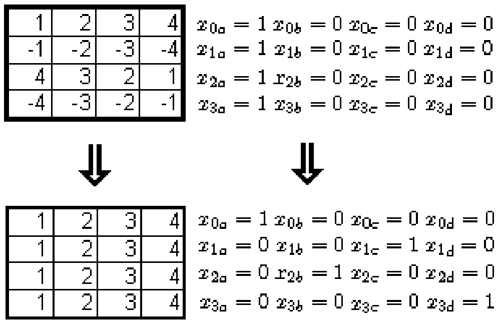


Fig. 2. Example of transitions of curvature matrix values and corresponding Boolean decision variables values. In details, row 1 is unchanged, row 2 has inverted sign only, row 3 has reversed order only, and row 4 has inverted both order and sign

the formulation of the optimization problem to a standard form. The first term of Equation 7 is constant w.r.t. the decision variables, so the problem can be formulated as a maximization problem over 0-1 variables:

$$\max_{x_{ia},x_{ib},x_{ic},x_{id}} \quad M = \sum_{j=0}^{v-1} \Big(\sum_{i=0}^{m-1} x_{ia} c_{ij} + \sum_{i=0}^{m-1} x_{ib} \overline{c}_{ij} - \sum_{i=0}^{m-1} x_{ic} c_{ij} - \sum_{i=0}^{m-1} x_{id} \overline{c}_{ij}\Big)^2 \tag{8}$$

where:

$$\overline{c}_{ij} = c_{i(v-j)}$$

and

$$x_{ra} + x_{rb} + x_{rc} + x_{rd} = 1 \ \ \forall\, r = 0 \ldots (v-1)$$

which implies that only one of the Boolean decision variables x_{ra}, x_{rb}, x_{rc} and x_{rd} must be true at once.

The problem can be re-formulated as an instance of a classical BQP (binary quadratic problem)

$$\max \quad x^T Q x = x^T V^T V x \tag{9}$$

with the additional constraint:

$$Ax = 1 \tag{10}$$

This translates into the fact that every row of the matrix can be moved in just one way, but different rows can be subject to different manipulations. The matrix V in Equation 9 is defined as:

$$V = (C, \overline{C}, -C, -\overline{C})$$

$v \times 4m$ matrix, where $\overline{C}$ is a $m \times v$ matrix whose elements are:

$$\overline{c}_{ij} = c_{i(v-j)}$$

The x^T vector is defined as

$$x^T = (x_a^T, x_b^T, x_c^T, x_d^T)$$

binary vector of dimension $4m$, while A is defined as

$$A = (I, I, I, I)$$

where I is the identity matrix of rank m. Finally, 1 in Equation 10 is the unitary vector of dimension m.

BQP is a well-know NP-Hard problem (Non-deterministic Polynomial-time Hard) [33], so we have implemented and compared three heuristic approaches for its solution.

The first heuristic approach is an ad-hoc Greedy algorithm in which the best of the four possible orientations was chosen for each row with a hill-climbing optimization heuristic that computes the objective function increase (or decrease) at every step. The classical problem in hill-climbing optimization is the pitfall of a local-minimum stop. In order to avoid this non-optimal solutions we have tested the results achieved by the Greedy algorithm against general purpose techniques.

- **Greedy approach**

Starting from matrix C the algorithm chooses for each row the orientation that yields the best objective function improvement. We define ΔM_{LR} as the variation of the objective function when we reverse the order of elements of row r, in symbols from step t to step $t+1$:

$$\begin{aligned}
&\Delta M_{LR} = M^{(t+1)} - M^{(t)} = \\
&= \tfrac{-2}{vm(m-1)} \textstyle\sum_{j=0}^{v-1} [(c_{rj}^{(t+1)} - c_{rj}^{(t)})(m\overline{c}_j^{(t)} - c_{rj}^{(t)})] = \\
&= \tfrac{-2}{vm(m-1)} \textstyle\sum_{j=0}^{v-1} [(c_{r(v-j)}^{(t)} - c_{rj}^{(t)})(m\overline{c}_j^{(t)} - c_{rj}^{(t)})]
\end{aligned} \tag{11}$$

and we define ΔM_{UP} as the variation of the objective function when we invert the sign of elements of row r, in symbols:

$$\begin{aligned} &\Delta M_{UP} = M^{(t+1)} - M^{(t)} = \\ &= \tfrac{-2}{vm(m-1)} \textstyle\sum_{j=0}^{v-1} [(c_{rj}^{(t+1)} - c_{rj}^{(t)})(m\overline{c}_j^{(t)} - c_{rj}^{(t)})] = \\ &= \tfrac{-2}{vm(m-1)} \textstyle\sum_{j=0}^{v-1} [-2c_{rj}^{(t)}(m\overline{c}_j^{(t)} - c_{rj}^{(t)})] \end{aligned} \tag{12}$$

and we define ΔM_{DG} as the variation of the objective function when we invert the sign of elements and reverse the curvature order of row r, in symbols:

$$\begin{aligned} &\Delta M_{DG} = M^{(t+1)} - M^{(t)} = \\ &= \tfrac{-2}{vm(m-1)} \textstyle\sum_{j=0}^{v-1} [(c_{rj}^{(t+1)} - c_{rj}^{(t)})(m\overline{c}_j^{(t)} - c_{rj}^{(t)})] = \\ &= \tfrac{-2}{vm(m-1)} \textstyle\sum_{j=0}^{v-1} [(-c_{r(v-j)}^{(t)} - c_{rj}^{(t)})(m\overline{c}_j^{(t)} - c_{rj}^{(t)})] \end{aligned} \tag{13}$$

where $\overline{c}_j^{(t)}$ is the mean on the jth column of the element of $C^{(t)}$ that can be updated from $C^{(t)}$ to $C^{(t+1)}$ with

$$\begin{aligned} \overline{c}_j^{(t+1)} &= \overline{c}_j^{(t)} + \tfrac{c_{rj}^{(t+1)} - c_{rj}^{(t)}}{m} = \\ &= \overline{c}_j^{(t)} + \tfrac{c_{r(v-j)}^{(t)} - c_{rj}^{(t)}}{m} \end{aligned} \tag{14}$$

for Equation 11,

$$\begin{aligned} \overline{c}_j^{(t+1)} &= \overline{c}_j^{(t)} + \tfrac{c_{rj}^{(t+1)} - c_{rj}^{(t)}}{m} = \\ &= \overline{c}_j^{(t)} + \tfrac{-2c_{rj}^{(t)}}{m} \end{aligned} \tag{15}$$

for Equation 12 and

$$\begin{aligned} \overline{c}_j^{(t+1)} &= \overline{c}_j^{(t)} + \tfrac{c_{rj}^{(t+1)} - c_{rj}^{(t)}}{m} = \\ &= \overline{c}_j^{(t)} + \tfrac{-c_{r(v-j)}^{(t)} - c_{rj}^{(t)}}{m} \end{aligned} \tag{16}$$

for Equation 13.

The Greedy algorithm stops when all rows have been considered and there is no further improvement. For each row the algorithm chooses the best improvement possible and iterates until all rows are in the optimal state, i.e. when no row change can lead to a $\Delta M < 0$. Details of the greedy implementation are reported in the pseudo-code below.

Code 1 Greedy Algorithm

INPUT: $C\_matrix^{(i)}$ { initial curvature matrix }
OUTPUT: $C\_matrix^{(f)}$ { optimal curvature matrix }
for all c column of C_matrix **do** compute $means(c)$
repeat
  for all r row of C_matrix **do**
    $\Delta M(r) = \min\{\Delta M_{LR}(r), \Delta M_{UP}(r), \Delta M_{DG}(r)\}$ { eq. 11, eq.12, eq.13 }
    if $\Delta M(r) < 0$ **then**
      if $\Delta M_{LR}(r) = \Delta M(r)$ **then**
        reverse order of row r
        for all c column of C_matrix **do** upgrade $means(c)$ { eq. 14 }
      else if $\Delta M_{UP}(r) = \Delta M(r)$ **then**
        reverse sign of row r
        for all c column of C_matrix **do** upgrade $means(c)$ { eq. 15 }
      else if $\Delta M_{DG}(r) = \Delta M(r)$ **then**
        reverse both sign and order of row r
        for all c column of C_matrix **do** upgrade $means(c)$ { eq. 16 }
      end if
    end if
  end for
until no row in C_matrix lead to $\Delta M < 0$

The algorithm is greedy because at every step the objective function increases. The ΔM can be calculated in linear time with respect to the number of columns v, so the convergence rate toward the optimal state is very fast.

- **Simulated Annealing**

Simulated annealing is a well-known technique used to approximate the solution of very large combinatorial optimization problems [34] and also to solve non linear optimizations in biomedical research, as in the image analysis of cancerous tissues [35] or in pattern recognition for electroencephalographic signals [36].

Each iteration of the algorithm consists of a random configuration move and the computation of the corresponding objective function variation. The moves that lead to an improvement are always accepted, while the others are accepted with the Boltzmann probability distribution.

$$P = exp^{-\frac{\Delta E}{kT}} \tag{17}$$

depends on the energy variations and on the temperature. Using our curvature-matrix C

$$\Delta E = \frac{\Delta M}{H} \tag{18}$$

is the objective function improvement normalized by

$$H = \frac{1}{v(m-1)} \sum_{j=0}^{v-1} \sum_{i=0}^{m-1} c_{ij}^2 \tag{19}$$

where k is the Boltzmann constant and T is the temperature parameter that gradually decreases during the run of the algorithm.

- **Genetic Algorithm**

Another general approach for solving combinatorial problem is genetic search [37], which has found wide application in practice [38].

A solution can be represented as a bit string of length $2m$ (for a search space of dimension 4^m), that describes the genome configuration for a genetic encoding of the problem. Starting from a randomly generated population of chromosomes each of them representing a different possible solution, the algorithm proceeds by selecting the better portion of the current population and applying random mutation and crossover, in order to select the best solution in every generation.

In Section IV-B we evaluated the performance of each approach and the minimum reached.

IV. Experimental Results and Discussion

In this section we report experimental results obtained by processing both real AFM images and computer-generated benchmark images generated according to Gaussian probability distribution as described in [22] so as to exhibit similar distribution and similar shape as in the real AFM images of DNA molecules (including tip-broadening effects).

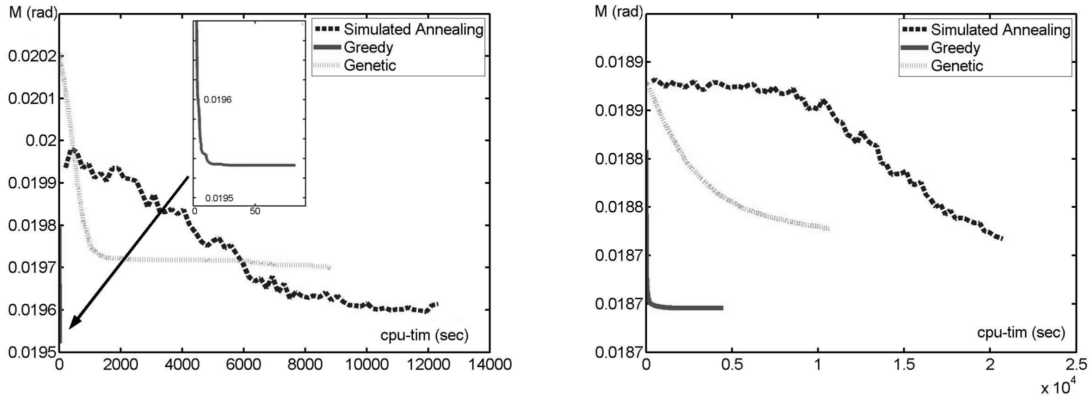
As AFM images we employed three sets of images characterized by molecules with different curvature profiles. The first and the second one were sets of images of palindromic dimers of DNA with two symmetric peaks of curvature while the third one was a set of images of a linear DNA molecule containing the highly curved kinetoplast DNA of the (Trypanosomatidae protozoan) *Crithidia fasciculata* characterized by a central high-curvature region.

As computer-generated images we employed two sets of images characterized by two different curvature profiles a priori chosen to be very similar to those of real molecules. The curvature profile has been chosen with a central peak of 0.06 rads in the first set and of 0.02 rads in the second one. Then, Gaussian random noise with variance 0.02 was added as a thermal perturbation. In addition, the molecules were randomly oriented.

A. Experimental Set Up

1) Length determination algorithm: To set the k parameter that minimizes the error in the the length calculation, the length determination algorithm was tested with computer-generated square images.

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. One thousand fragments were employed for the seven fragment sizes and the three additive noises. As a result, we found two different values 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).



(a) (b)

Fig. 3. Performance of heuristic optimization approaches. (a): Results on high-curvature simulated images. Greedy optimal value is $19.52 \times 10^{-3} rad^2$ with a run time of $82.24 sec$. Simulated annealing runs in a range of temperature $[10 - 0.15]$ with a reduction rate of 0.98 every 4000 moves. Genetic algorithm runs for 4000 generations; (b): results on low-curvature simulated images. Greedy optimal value is $18.69 \times 10^{-3} rad^2$ with a run time of $4481 sec$

Finally, we chose for the 3.9 nm pixel size images k_{opt}=0.32 and for the 7.8 nm pixel size images k_{opt}=0.16. Details about these tests and settings are reported in[26].

2) Fragment flipping algorithm: The performance of the optimization methods of Fragment flipping algorithm have been tested with three different data-sets of molecules for each heuristic approach. The first two sets are 445×270 and 1008×280 curvature matrices obtained by the two computer-generated set of images, while the third was a 1008×280 matrix obtained by the set of molecules containing the kinetoplast DNA of *Crithidia fasciculata*. The simulated annealing implementation has been set up with a temperature that decreases in the range $[10.0 - 0.15]$ with a reduction rate of 0.98 every 4000 moves. In the genetic optimization algorithm we used 40 individuals per population, a selectivity rate of 0.9 with a recombination rate of 0.7 using a single-point crossover, and a single mutation rate of $0.7/2n$ where $2n$ is chromosome length. The evolution run for 4000 generations.

B. Performance of Heuristic Approaches

1) Comparison between Greedy, Simulated Annealing and Genetic Algorithm: In this section we report experimental results on performance evaluation obtained by running the three heuristic approaches, greedy, simulated annealing and genetic, on the two sets of computer generated benchmarks images and on real AFM images in order to evaluate the best one for the intrinsic curvature profile reconstruction problem.

Subsequently, we show results of the comparison between the performance of the best heuristic approach and an exhaustive approach in order to evaluate the distance between the minimum reached by our heuristic algorithm and the global minimum reached by the exhaustive search method. The exhaustive approach is the exhaustive evaluation of all the molecular orientations in order to find the best one that minimize the M function (see Equation 5).

Figures 3 and 4 show the quality of the results of the three optimization methods with respect to cpu-time on computer-generated data and real AFM data, respectively. Figure 3.a shows the results of the three algorithms with high-curvature simulated images. The greedy implementation reaches a minimum of $19.52 \times 10^{-3} rad^2$ that is the best value, while $19.59 \times 10^{-3} rad^2$ and $19.69 \times 10^{-3} rad^2$ are the values respectively of simulated annealing and genetic implementation. The plot evidences that greedy implementation is the fastest, with a very short running-time, less than 2 minutes versus the 2-4 hours of the other methods.

With the lower-curvature simulated images the best approach is, once again, the greedy (see figure 3.b). The minimum value reached is $18.69 \times 10^{-3} rad^2$, while $18.77 \times 10^{-3} rad^2$ and $18.78 \times 10^{-3} rad^2$ are the values respectively of simulated annealing and genetic implementation. Greedy running-time is about five times shorter than the simulated annealing result and three times shorter than the genetic one.

For the set of experiments on real AFM data we employed images of a linear molecule containing the curved *Crithidia fasciculata* fragment. As result, the three heuristic methods maintain the same behavior observed in the computer-generated benchmark cases, as shown in Figure 4. The Greedy implementation reaches a minimum of $22.79 \times 10^{-3} rad^2$, the simulated annealing of $22.94 \times 10^{-3} rad^2$ and the genetic of $23.22 \times 10^{-3} rad^2$. The time performance of the Greedy optimization is once again the best one, only $106 sec$ for reaching the minimum. The comparison between the optimization methods shows that the hill-climbing approach is the best both with respect to the

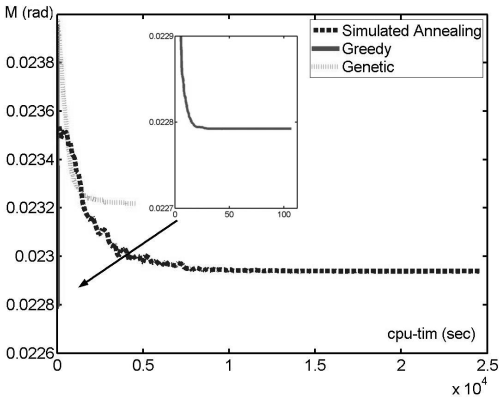


Fig. 4. Performance of heuristic optimization approaches on real molecules images. The Greedy implementation achieved the best performance of $22.79 \times 10^{-3} rad^2$ with a run time of $106.6 sec$

objective function optimization and with respect to the run-time. In fact, comparing the intrinsic curvature profiles reconstructed by using the greedy approach, the simulated annealing and the genetic algorithm, we found that the greedy implementation is more able than the two other heuristic methods to reconstruct the curvature profile and to detect the intensity of the peaks with an error of one order of magnitude smaller than errors provided by the two other methods. Details about greedy performance in the detection and reconstruction of the intrinsic curvature profile are reported in the Sec. IV-D.2.

2) Greedy versus Exhaustive Search Solution: In this subsection we report experimental results on the accuracy of our greedy algorithm with respect to the exact solution of the mathematical problem in Equation 5, found by an exhaustive search.

A polynomial algorithm that finds the exact solution of the BQP problem (i.e. a NP-hard problem) is not known. Thus, to evaluate the accuracy of our method we have compared our results with the minimum obtained by exhaustive search in the space of allowed configurations. This search is considerably time-consuming. In fact the execution time grows exponentially with the number of the molecules, that is the number of the rows in the curvature matrix. Thus, to assess the deviation from optimality as a function of problem dimension we tested our greedy optimization approach by using several matrices with increasing number of the rows up to a maximum of 15 rows. For each matrix we computed the global optimum by exhaustive search and we computed 10 greedy solutions starting from different random row orientations.

Figure 5.a shows the comparison between the exhaustive and the greedy solutions. The coordinates of each marked point in the figure represent the optimum value (on the x axis) and the greedy solution value (on the y axis). Different symbols are used for different matrix sizes. The bisector line corresponds to the ideal condition: the greedy solution has the same value of the optimum. It can be observed that there is a strong correlation between the greedy solutions and the exhaustive ones. Notice also that the quality of the solutions are not dependent on the problem size. The correlation coefficient of the regression line is 0.9981.

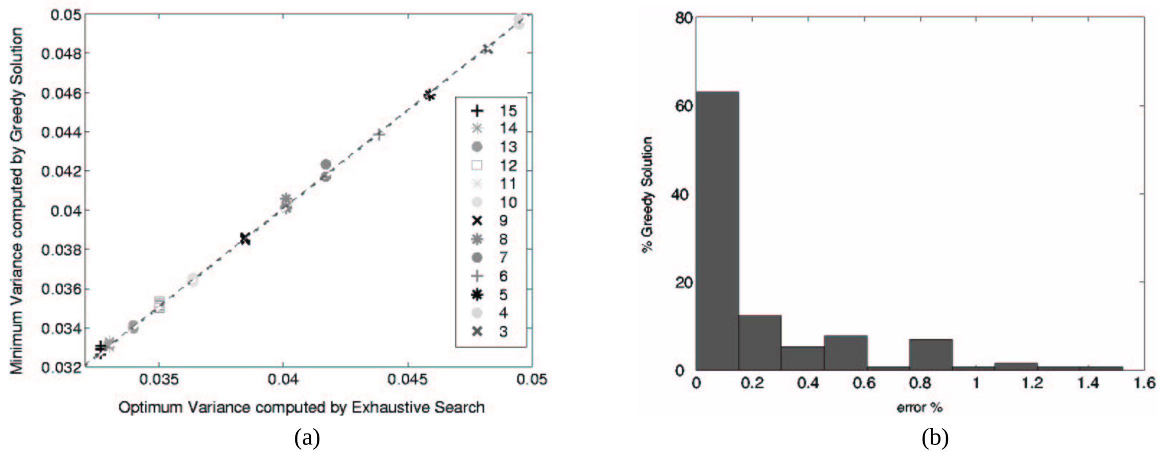


Fig. 5. (a): Correlation between greedy solution and global optimum. The dashed line is the regression line; (b): Distribution of the percentage error of the greedy algorithm with respect to the global minimum. The global optimum computed by an exhaustive search and the greedy optimization lead to very close results. More than 97% of the greedy solutions reach the minimum with a deviation from global optimum smaller than 1%

In the Figure 5.b we reported the distribution of percentage error of the greedy solutions with respect to the global optimum. For every matrix employed for the tests, at least one of the 10 greedy solutions achieves the global optimum. The 97% of greedy solutions are 1% deviation from the global optimum and the 63% are smaller than 0.2% deviation from the global optimum. Overall, the greedy optimization find a minimum with an average error of 0.18% of global optimum.

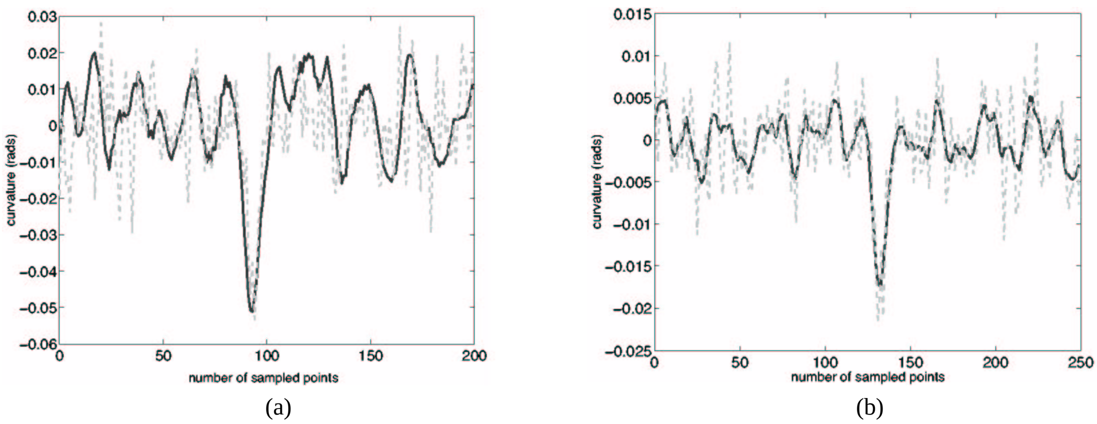
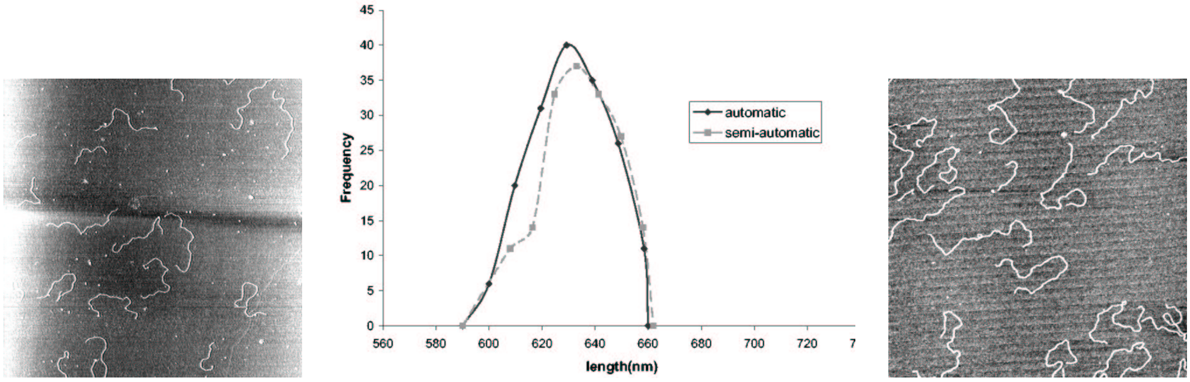


Fig. 6. Original (dashed plot) and reconstructed profile of curvature (solid plot) for simulated DNA-molecules. (a): molecules with a high curved region with one peak of 0.06 rads; (b): molecules with a central curved region with one peak of 0.02 rads



(a) (b) (c)

Fig. 7. (a): EcoRV-EcoRV dimer DNA molecules in a 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; (c) Molecules containing the kinetoplast DNA of *Crithidia fasciculata* in an AFM image.

C. Curvature computation on Simulated Images

As already introduced, in the first set of computer generated benchmarks the chosen curvature profile has a high curved region with one peak of 0.06 rads. As result of our ad-hoc greedy approach, the algorithm provides a very high percentage of molecular-orientation detection: 96.79 % from a population of 655 molecules. Regarding curvature estimation, in Figure 6.a we report the reconstructed intrinsic curvature profile (solid curve) compared with the original curvature profile (dashed curve). This comparison shows that the estimated profile well approximates the original one, pointing out the high estimation capability of our algorithm. The reconstructed profile matches the original one with a deviation of 10.9nm, that is 1% of molecule length. Moreover the magnitude of the peak was found with respect to the original one with an error of $6.79 \cdot 10^{-4}$ rads.

We repeated the same experiments for the second test set. In this case the chosen curvature profile has one peak of 0.02 rads. As a result, our algorithm provides 76.19 % of molecular-orientation detection from a population of 902 molecules. Figure 6.b reports the reconstructed intrinsic curvature profile (solid curve) compared with the original profile (dashed curve). Also in this case our algorithm shows significant estimation capability. The reconstructed profile matches the original one with a deviation of 8.7nm, that is 0.8% of molecule length. Moreover the magnitude of the peak was found with respect to the original one with an error of $8.3 \cdot 10^{-3}$ rads.

D. Results on Real Images

Experimental results on real AFM images regard automated curvature estimation of three DNA molecules: a linear molecule containing the curved *Crithidia fasciculata* fragment, and two palindromic dimers (EcoRV-EcoRV and PstI-PstI) [15] [18].

1) Length computation on Real Images: Before evaluating the curvature profiles DNA length distributions should be computed for each data set to verify that significant DNA structural transitions have not occurred in the DNA deposition process. This test is necessary to verify the validity of the data set for the curvature profiles calculation.

We computed length distributions on the three sets of real images employed in the curvature profile estimate. All sets were images at resolution of 3.9nm/pixel, so k parameter in length computation procedure was set to 0.32.

For each data set, we compared our results with the expected real length value to verify that no significant structural transitions have occurred in the deposition process. In fact a B-to-A transition in the secondary DNA structure involves in the molecules a shortening of about 18% - 30%. We performed also a comparison with length distribution obtained with a semi-automated method [22][18] widely used in bio-labs for its suitable accuracy. In fact, the semi-automated procedure can be very effective for selecting molecules of interest because of the ability of the human-eye to distinguish molecules from background noise or artifacts.

The first and second data sets displayed 633.4nm palindromic dimers of DNA obtained by joining two segments in either the head-to-head or tail-to-tail configuration [18]. The molecules were cut between the EcoRV and the PstI sites and dimerized around either site to get a EcoRV-EcoRV dimer and a PstI-PstI dimer. Figure 7.a shows an example of the first dimer set of images. The length distribution of the molecules in the first data set is shown in Figure 7.b (solid curve). Observing the length distribution histogram in this figure 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 (dashed curve). 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 an average length of 631.3nm with a standard deviation of 14.7nm. The average length is thus very close to the expected one (633.4nm) with a 0.33% error. Moreover, comparing our results to the semi-automated measures we have observed a deviation about 0.57%, as seen in Figure 7.b. From these results, we concluded that no significant structural transitions seem to have occurred in the deposition process. A similar result was obtained for the second dimer data set (i.e a mean value of distribution of 635.37nm, that is 0.3% of the expected length and a standard deviation of 14.68nm).

The third set of images displayed the 1098nm DNA molecules containing the kinetoplast DNA of the (Trypanosomatidae protozoan) *Crithidia fasciculata*. Figure 7.c shows an example of these set of images. The molecules appear in the image characterized by very irregular profiles due to the high molecule curvature. This property translates into a harder computation of molecule lengths and profiles because surrounding noise shadows DNA shapes proportionally to DNA profile complexity. We obtained an average length of 1085nm with a standard deviation of 20.5nm. The average length is thus very close to the expected one (1098nm) with a 1.18% error and the results are comparable with the first and the second set of molecules. With respect to the semi-automated measures we have observed a deviation of about 0.77%.

Also with this data set, we can conclude that no significant B-DNA to A-DNA transitions seem to have occurred during the AFM sample preparation. This verifies the validity of the data sets for the intrinsic curvature profile computation.

2) Curvature analysis on Real Images: As introduced in previous subsections, experimental results on real AFM images regard automated curvature estimation of the three sets of DNA molecules. The results are compared with theoretical profiles obtained using a referenced theoretical model for predicting sequence-dependent curvature [39] [40] [21] [41].

The molecules containing the kinetoplast DNA of *Crithidia fasciculata* are characterized by a high central curved region with a theoretical peak of 0.196 rads [20]. The number of curvature sample points was set to 250 in order to obtain about 13 bp (base pair) per segment, the highest density of data points that are meaningful considering the micrograph resolution.

Our goal is to detect the location of the peak of curvature along the molecule and to reconstruct the curvature profile, in particular in the area of the peak. That is most relevant for curvature effects in biological studies. Figure 8 reports the reconstructed intrinsic curvature

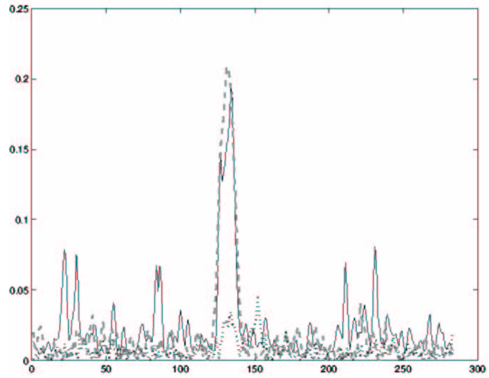


Fig. 8. Theoretical (solid plot), reconstructed curvature (dashed plot) and reconstructed curvature profile without orientation detection (dotted plot). The number of curvature sample points was set to 250 in order to obtain about 13bp per segment, the highest density of data points that are meaningful considering the micrograph resolution

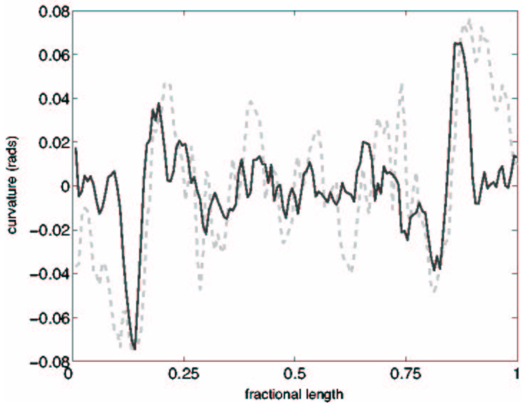
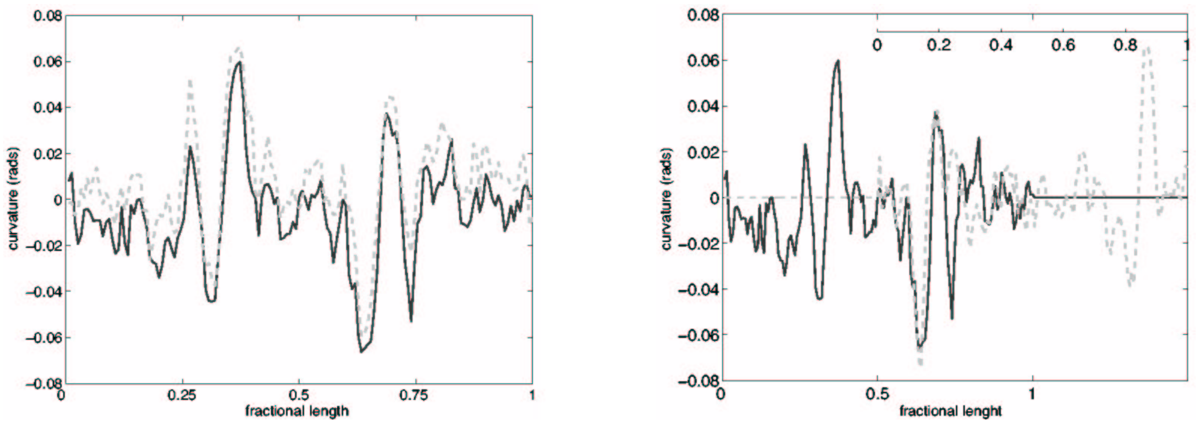


Fig. 9. Theoretical (dashed plot) and EcoRV-EcoRV reconstructed intrinsic curvature profile (solid plot). The number of curvature sample points was set to 148 in order to obtain about 13bp per segment, the highest density of data points that are meaningful considering the micrograph resolution

profile using our greedy approach (dashed curve) compared with the theoretical curvature module profile (solid curve) and the curvature profile reconstructed without orientation detection of the molecules in the image (dotted curve). The comparison shows how the orientation detection enhances the curvature profile estimation. From the plot it comes in evidence that reconstructed intrinsic curvature profile well approximates the theoretical one with a mean square error of $3.8122 \cdot 10^{-4}$ rads in the region of the peak. Moreover, the algorithm correctly detects the location of the peak in the molecule with a deviation w.r.t. the theoretical one of 14.1 nm, that is the 1.28% of molecule length.

EcoRV-EcoRV dimer molecules are characterized by two curvature symmetric peaks with a theoretical value of about 0.08 rads[40]. We set the number of curvature sample points to 148 for both EcoRV and PstI dimer sets in order to obtain about 13bp per segment, that is the highest density of data points that are meaningful considering the micrograph resolution. For the same reason, the theoretical curvature profile[15] was calculated with a similar number of curvature sampled points. As expected, the significant peaks of curvature for EcoRV-EcoRV dimer are found near the ends of the molecules and the location of the peaks matches the theoretical one very well with a deviation of 8.44nm, that is 1.3% of molecule length. In Figure 9 we report the reconstructed intrinsic curvature profile (solid curve) compared with the theoretical curvature module profile (dashed curve). From the plot it comes in evidence that reconstructed intrinsic curvature profile well approximates the theoretical one with a standard deviation in the regions of the peaks of $6.1 \cdot 10^{-3}$ rads.

PstI-PstI dimer molecules are instead characterized by some significant curvatures in the central part of the molecules. Also the PstI-PstI molecules are characterized by symmetric curvature peaks with a theoretical value of 0.08 rads. The number of curvature sample points was set to 148 in order to obtain about 13bp per segment. In Figure 10.a we represent the PstI-PstI reconstructed intrinsic curvature profile (solid



(a) (b)

Fig. 10. (a): PstI-PstI intrinsic curvature profile (solid plot) compared with the same data with the sign flipped and reversed (dashed plot) to evidence the symmetry of the computed profile; (b): PstI-PstI intrinsic curvature profile (solid plot) compared with the EcoRV-EcoRV reconstructed intrinsic curvature profile (dashed plot)

curve) compared with the same data (dashed curve) with the sign flipped and reversed to evidence the symmetry of the computed profile. Moreover, in the PstI-PstI curvature profile computation we expected to find high curved regions as the EcoRV-EcoRV dimer molecules, in the first in the center and in the second in the boundaries.

In Figure 10.b we compare the PstI-PstI reconstructed intrinsic curvature profile (solid curve) with the EcoRV-EcoRV reconstructed intrinsic curvature profile (dashed curve) shifted of half molecule length to verify if the curvature regions overlap. As it can be noted in the figure, the two curvature profiles are characterized by exactly same peak positions along the molecule. Moreover, the PstI-PstI reconstructed intrinsic curvature profile well approximates the EcoRV-EcoRV one with a standard deviation of 0.0157 rads.

E. Discussion

From a biological perspective, we are interested on peak localization and profile. In fact, from peak position and intensity we derive important consequences related to biological processes. It is also worth noting that even if AFM has high resolution with respect to other methods, the pixel size of AFM images is an upper bound on the accuracy achievable by following image processing steps. Nevertheless the accuracy obtained is useful for today's biological analysis steps like DNA-protein interactions and DNA transcription.

Concerning peak profile reconstruction, the algorithm shows higher accuracy for higher curvature peaks, because in this case the detection of molecule orientations on the substrate is more effective. In particular, the standard deviation obtained on peak reconstruction of EcoR V dimer is of $6.1 * 10^{-3}$ rads. The accuracy achieved allows to perform analysis on DNA molecules involved in DNA-protein interactions or other biologically relevant functions. In fact, molecules that show a difference in the peak curvature profile lower than this value do not show significant differences from a biological perspective. Moreover, the intensity of curvature peaks of these molecules is quite common. This means that our algorithm performs well on a wide range of cases. For molecules characterized by a higher peak curvature level the achieved accuracy is even higher, as shown in the case of molecules containing the kinetoplast DNA of Crithidia fasciculata.

As regards localization, we showed that the proposed algorithm gives an error around 1% w.r.t. the theoretical profile in peak localization. Also in this case, the upper bound on the accuracy is given by the error due to the micrograph resolution.

These accuracy results have been obtained with a fully automated method that allows to consistently speed up the analysis. Moreover, our automated method allows detecting the orientation of molecules on the substrate without requiring tags and it allows to reconstruct DNA intrinsic curvature profiles avoiding the production of palindromic constructs (dimers) of molecules under analysis which is required to overcome the uncertainty of molecular orientation.

V. Conclusions

In this paper we have presented a fully automated algorithm for DNA intrinsic curvature profile computation in Atomic Force Microscope images through the detection of the correct spatial orientation of the molecules on the AFM substrate following the DNA deposition process.

We proposed a fast heuristic orientation finding algorithm, that modifies one molecular orientation at a time with linear-time heuristic transitions. We performed extensive comparisons between different heuristics, and we found that our ad-hoc greedy implementation is the best both with respect to the objective function optimization and with respect to the run-time. We then compared the results obtained with our method against an exhaustive technique (which is obviously impractical when dealing with a large number of molecules) to show the effectiveness of our approach. We found that there is a strong correlation between the greedy results and the exhaustive ones.

The algorithm detects fragment orientation on AFM images without labels with a percentage of correct molecular-orientation detection of 96.79% in computer-generated benchmarks, for molecules with a high curvature peak.

To test the capability of our method to reconstruct the intrinsic curvature of real DNA molecules, different sets of DNA images were obtained by deposition of two different DNA dimers and of kinetoplast DNA of the *Crithidia fasciculata*. The length determination results show that no significant B-to-A transition in the DNA secondary structure have occurred in the deposition process. This verifies the validity of the real data sets for the intrinsic curvature profile computation. Then, the comparison of our results on these sets of real data with those obtained with a theoretical approach referenced by several papers in the field, shows a negligible mean square error in the curvature profile estimation: $3.8122 \cdot 10^{-4}$ rads over a profile with a central peak value of 0.196 rads, and $6.1 \cdot 10^{-3}$ rads over a curvature profile with two symmetric peaks of about 0.08 rads. Moreover, it shows an error of about 1% in the correct detection of the location of the peaks along the molecules.

VI. Acknowledgments

G.Z. and B.S. wish to acknowledge support from *Progetti Pluriennali Universita' di Bologna*; FISR D.M. 16/10/20 year 1999 and the ESF Eurocore SONS program (2003/2006).

References

[1] Jauregui R, Abreu-Goodger C, Moreno-Hagelsieb G, Collado-Vides J, Merino E "Conservation of DNA curvature signals in regulatory regions of prokaryotic genes" *Nucleic Acids Res* 2003 Dec 1; 31(23):6770-7

[2] Travers, A. Muskhelishvili G. “DNA microloops and microdomains: a general mechanism for transcription activation by torsional transmission” *J Mol Biol* 1998 Jun 26; 279(5):1027-43

[3] Travers, A. A “DNA-Protein Interactions” (Chapman and Hall, London), 1993

[4] Dickerson, R. E “Sequence-dependent helix deformability in the recognition of B-DNA”, *Biopolymers*, 1997, 44, 321

[5] Dickerson, R. E, Chiu, T. K “Helix bending as a factor in protein/DNA recognition”, *Biopolymers*, 1997, 44, 361-403

[6] Muzard G., Theveny B., Revet B. “Electron microscopy mapping of pBR322 DNA curvature. Comparison with theoretical models”, *EMBO J*, 1990, 9, 1289-1298

[7] Bednar J., Furrer P., Katritch V., Stasiak A. Z., Dubochet J., Stasiak A. “Determination of DNA persistence length by cryo-electron microscopy. Separation of the static and dynamic contributions to the apparent persistence length of DNA”, *Mol. Biol.*, 1995, 254, 579-594

[8] C. Rivetti, M. Guthold, C. Bustamante “Scanning force microscopy of DNA deposited onto mica: equilibration versus kinetic trapping studied by statistical polymer chain analysis”, *Journal of Molecular Biology*, 1996, Vol.264 (5), 919-32

[9] C. Rivetti, C. Walker, C. Bustamante “Polymer Chain Statistics and Conformational Analysis of DNA Molecules with Bends or Sections of Different Flexibility”, *J. Mol. Biol.*, 1998, Vol.280, 41-59

[10] Crothers D. M., Drak J., Kahn J. D., Levene S. D. “DNA curvature and deformation in protein-DNA complexes: A step in the right direction”, *Methods Enzymol.*, 1992, 212, 3-29

[11] Shore D., Baldwin R. L. “Energetics of DNA twisting. II. Topoisomer analysis”, *J. Mol. Biol.*, 1983, 170, 983-1007

[12] Haran T. E., J. D. Kahn, et al. “Sequence elements responsible for DNA curvature” *Journal of Molecular Biology*, 1994, 244 (2):135-43

[13] Harvey S. C., M. Dlakic, et al. “What is the basis of sequence-directed curvature in DNAs containing A tracts?” *Journal of Biomolecular Structure and Dynamics*, 1995, 13 (2):301-7.

[14] Cognet J. A., Pakleza C., Cherny D., Delain E., Cam E. L. “Static curvature and flexibility measurements of DNA with microscopy. A simple renormalization method, its assessment by experiment and simulation”, *J Mol Biol*, 1999, 285 (3): 997-1009

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

[16] A. Scipioni, C. Anselmi, G. zuccheri, B. Samorı, P. De Santis, “Sequence-dependent DNA curvature and flexibility from scanning force microscopy images”, *Biophys. J.*, 2002, 83(5):2408-18

[17] 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

[18] G. Zuccheri, B. Samorı, “Scanning Force Microscopy Studies on the Structure and Dynamics of Single DNA molecules” *Atomic Force Microscopy in Cell Biology*, Academic Press, 2002, 68(17):357-95

[19] B. Theveny, B. Revet, “DNA orientation using specific avidin-ferritin biotin end labelling”, *Nucleic Acids Res.* (1987) 15: 947.

[20] P. De Santis, A Palleschi, M. Savino, A. Scipioni, “A theoretical model of DNA curvature”, *Biophys. Chem.*, 32 (1988) 305-317

[21] C. Anselmi, G. Bocchinfuso, P. De Santis, M. Savino, A. Scipioni, “Dual role of DNA intrinsic curvature and flexibility in determining nucleosome stability”, *J. Mol. Biol.*, 286 (1999) 1293-1301

[22] 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

[23] I. Muzzalupo, C. Nigro, G. Zuccheri, B. Samorı, C. Quagliariello, M. Buttinelli “Deposition on mica and scanning force microscopy imaging of DNA molecules whose original B structure is retained” *Journal of Vacuum Science and Technology A (Vacuum, Surfaces, and Films*, 13 (3,pt.2):1752-4, 1995

[24] C. Bustamante, J. Vesenka, C. L. Tang, W. Rees, M. Guthold, R. Keller “Circular DNA molecules imaged in air by scanning force microscopy”, *Biochemistry*, 1992, 31, 1, 22-26

[25] 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

[26] E. Ficarra, L. Benini, E. Macii, G. Zuccheri "A Robust Algorithm for Automated Analysis of DNA Molecules in AFM Images" *IASTED Biomedical Engineering (BioMED 2004)* Innsbruck, Austria, Feb.2004

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

[28] 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

[29] 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

[30] 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

[31] Landau, L.D. and Lifshitz, E.M. "Theory of Elasticity", *Pergamon Press*, 1986, Oxford, NY

[32] H. W. Eves, "A Survey of Geometry, rev. ed.", *Allyn and Bacon* Boston, MA, pp.256 and 262, 1972.

[33] M. R. Garey, D. S. Johnson. "Computers and intractability: A guide to the theory of NP-completeness" *W.H. Freeman and Co*, 1979.

[34] S. Kirkpatrick, C. D. Gelatt, Jr., M. P. Vecchi "Optimization by Simulated Annealing" *Science*, Nr. 4598, 13 May 1983

[35] K. Jafari-Khouzani, H. Soltanian-Zadeh, "Multiwavelet grading of pathological images of prostate" *Biomedical Engineering, IEEE Transactions on*, Vol.50 (6):697-704, June 2003

[36] D. Khosla, M. Singh, M. Don, "Spatio-temporal EEG source localization using simulated annealing" *Biomedical Engineering, IEEE Transactions on*, Vol. 44:716-723, June 1997.

[37] J.Holland. "Adaptation in Natural and Artificial Systems" *University of Michigan Press*, Ann Arbor, USA, 1975.

[38] A. Gacek, W. Pedrycz, "A genetic segmentation of ECG signals" *Biomedical Engineering, IEEE Transactions on*, Vol.50 (10):1203-1208, Oct. 2003

[39] P. De Santis, A Palleschi, M. Savino, A. Scipioni, "Validity of the nearest-neighbor approximation in the evaluation of the electrophoretic manifestations of DNA curvature" *Biochemistry*, 29 (1990) 9269-9273

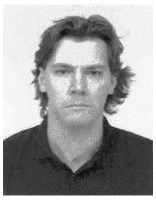
[40] P. De Santis et al., *Welcome to WebDNA*, http://archimede.chem.uniroma1.it/webdna.html

[41] C. Anselmi, G. Bocchinfuso, P. De Santis, M. Savino, A. Scipioni, "A theoretical model for the prediction of sequence-dependent nucleosome thermodynamic stability", *Biophys. J.*, 79 (2000) 601-613

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 of Torino under the supervision of Prof. Enrico Macii. From 2001 she is also research assistant in the microelectronics laboratory at the department of electronics and computer science in the University of Bologna under the supervision of Prof. Luca Benini. In 2002 she was visiting at the Computer Systems Laboratory (CSL), Stanford University, CA. Now she is intern at the EPFL de Lausanne, Faculte de Informatique et Comunications. Elisa Ficarra's research interests include computer vision, biomedical and molecular imaging, gene expression analysis, gene clustering and networks.

Daniele Masotti Daniele Masotti is PhD student at DAUIN - Politecnico di Torino (Italy). In 2003 He received a Master's degree in bioinformatic from Turin University. He received a Computer Science Degree in 2001 from the University of Bologna (Italy). His research interests include combinatorial optimization algorithms for DNA structural analysis and statistical methods for microarray data analysis.

Luca Benini Luca Benini is an Associate Professor at the Department of Electrical Engineering and Computer Science (DEIS) of the University 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 Symposium 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).

Bruno Samor Bruno Samori' is Professor of Organic Chemistry at the School of Biotechnology of the University of Bologna. He carried out research with polarized spectroscopy techniques on liquid crystalline materials until 1990, at the University of Bologna, at King's College of London with S. F. Mason, at the University of California Berkeley with I. Tinoco and at the University of New Mexico with C. Bustamante. The focus of his research then moved to studies of DNA superstructures by using the Atomic Force Microscopy (AFM). His research activities are presently focused on DNA-based nano-bio-technologies and on single-molecule mechanochemistry in recognition and adhesion events in molecular and cell biology. Bruno Samor is a member of the Editorial Board of ChemBioChem (Wiley-VCH), and of the Scientific Advisory Committee of the International Society for Nanoscale Science, Computation and Engineering (ISNSCE). He has been President of the Division of the Chemistry of the Biological Systems of the Italian Chemical Society and also of the Italian Society of Microscopic Sciences.