Graphite-MicroMégas, a tool for DNA modeling

Samuel Hornus* INRIA
Damien Larivière† Fondation Fourmentin-Guilbert

1 3D modeling : An engine of hypotheses and new questions

One of the great challenges of biology is to map the relations between macromolecular components to provide models of the cellular processes. Nearly every major process in a cell is carried out by assemblies of ten and more proteins which interact with several other complexes. DNA is also a major determinant of the global cell organization. Its own spatial and dynamic organization due in part to its association with proteins is essential in numerous biological processes. It can be either highly compacted at the central region of the bacterial cells, wrapped around proteins in the nucleus of animal cells. In each situation, the way it is packaged directly influences the functioning and the fate of the cell [Vendeville et al. 2011].

However, fully assembled machineries or short-lived intermediates often have proved refractory to structure determination. DNA is also quite hard to detect with good resolution from current microscopic techniques.

Tools allowing to reconstruct protein machineries and DNA in 3D are needed to investigate their possible arrangements within the finite cellular volume. 3D modeling and visualization are thus called to play a leading role in the understanding of the cellular architecture.

A 3D modeling tool set allows scientists to generate workable 3D scenes made of proteins and nucleic acids spatially but hypothetically distributed according to biological data. The coupled processes of scene creation (also called scene composition) and real-time immersion within the newly-created biological arrangement is a valuable experience. The reason behind is that the user effectively deals with a finite number of objects (proteins and or nucleic acids) within a limited space. These objects often display a spatial pattern and interact in restricted ways. Only a few organizations of macromolecules put together can respect these constraints. The user has consequently the opportunity to identify such architectures and to draw testable hypotheses about the shape of an architecture, the number of elements involved within, or its size and then its function [Goodsell and Johnson 2007].

2 Presentation of MicroMégas

MicroMégas is the current state of an ongoing effort to develop such a set of tools. We here present its DNA modeling part. MicroMégas is implemented as a plug-in to Graphite, which is a research platform for computer graphics, 3D modeling and numerical geometry that is developed by members of the ALICE team of INRIA. [1] A webpage has been set up that contains a user manual and describes the installation procedure for MicroMégas. [2]

Apart from the DNA modeling and visualization tool, MicroMégas also features facilities for meshing a protein surface and coloring selected parts of it, as well as inter-residues distance calculation.

* samuel.hornus@inria.fr
† damien.lariviere@fourmentinguilbert.org

[1] `alice.loria.fr/index.php/software.html`

[2] `www.loria.fr/~shornus/FFG/micromegas.html`

2.1 The DNA modeling tool

The modeling typically starts with the generation of a protein surface, that will guide the modeling of the DNA strand that should interact with it. Our DNA modeling process is straightforward. We use standard tools to design a curve in space. Currently in MicroMégas, two types of curves can be edited: quadratic and cubic Bézier curves; furthermore, a sequence of points can be imported, which will create a smooth curve interpolating the sequence. For Bézier curves, control points can be created, moved, deleted and duplicated, making the modeling process easy.

At any time during the modeling session, the curve can be visualized differently, as a plain line, a tube or as DNA. The visualization can be made partially transparent so as to continue editing the curve while seeing the DNA model move and transform in real-time.

When the overall shape of the DNA has been decided, it is possible to fine tune the position of its base pairs: the user can force the angular position of any chosen base pair around the curve. The angular position of the remaining pairs is then automatically set so as to get as close as possible to the natural DNA helix. In this way, for example, DNA can be locally untwisted, as illustrated in Figure 5.

Finally, it is possible to export the modeled DNA strand in the PDB file format, provided that the total number of atoms is less than 10^5. This is useful for checking the physical integrity of the strand with energy minimizing tools, and for further processing of the strand with other softwares.

To our knowledge, MicroMégas provides the first interactive tool dedicated to DNA modeling, and is far easier to use than previous approaches using e.g. fully general 3D modeling software.

3 Example applications: the bacterial DNA repair process

Replication of the chromosome in bacteria is not error free: The error rate of the DNA polymerase reaches 1 over 10^6 synthesized base pairs [Kunkel 2004]. For the E.coli genome, it results in about 5 wrong pairings (like GT or AC pairing) over one replication round. The repair process is coordinated by the proteins MutS, MutL and MutH. The 3D structure of the complex they form with or without DNA is unknown. Structural models for the complex were generated based on the crystal structures of the individual proteins (see [Winkler et al. 2011] for details). Thanks to the Graphite-MicroMégas tool, it is possible to integrate atomic DNA within an arrangement (Figure 3) by simply drawing a trajectory going through several known or hypothetical binding sites at the surface of the proteins. The benefits are multiple as it allows to identify new binding sites and to design experiments like chemical crosslinking or FRET for a purpose of validation, or to discriminate between protein arrangements when the global DNA shape (straight or loop) is experimentally known.

Other examples Graphite-MicroMégas allows to study arrangements of nucleosomes [Davey et al. 2002] (Figure 4) and to explore the packaging of very long DNA portion [Junier et al. 2010] (Figure 1).

4 Implementation

This section give technical detail about the implementation of the DNA modeling tool.

We choose to use a smooth curve C to model DNA: at any point on the curve, the tangent vector must be well defined. So must be the arc-length from the start of the curve to that point.

We use two point samplings of C: an *adaptive sampling* of C consists of a sequence of point on C that form a linear approximation of the curve such that the angle between two consecutive segments is below some threshold. This adaptive sampling is used to display a visually smooth curve while minimizing the number of segments used in the linear approximation of the curve.

We also use an *uniform sampling* S_u of P, in which all sub-curves between consecutive sample points have the same arc-length. We use a spacing close to $3.4\,\text{Å}$ to obtain a unifom sampling S_u in which the samples can serve as anchor points for base-pairs of DNA.

4.1 A rigidly moving frame

From the above, we have obtained a uniform sampling S_u of the curve together with the tangent vector at every sample point. In order to coherently orient the base pairs along the curve, we need to augment this data with a normal vector, so as to obtain an orthonormal frame at each sample point. Many simple methods to do so result in a sequence of frames exhibiting discontinuities or strong torsion (eg, the Frenet frame). But ideally, we would like the sequence of frames to be as rigid as possible (without discontinuities and minimizing torsion) in order for it to be a good start for orienting the DNA base pairs. This is not an easy ask, but the work Wang et al. shows how one can obtain an extremely good approximation at a very low computational cost [Wang et al. 2008]. After computing a normal vector at each sample point of S_u in this way, we store the sequence of triple $\{f_i = (o_i, t_i, n_i), i \in [1, N]\}$ where t_i is the tangent vector and n_i is a normal vector to the curve C at the sample point o_i (Figure 2).

4.2 The DNA strand model

MicroMégas does not support a specific sequence of nucleotides yet, and thus assumes a generic one (`AAA...`). We keep in memory an atomic model of a base pair `AT` centered at the origin. Then, in order to model or draw the full DNA strand, we instanciate the atomic (3D) model at each sampling point o_i of S_u in the frame f_i. To reproduce the DNA helix, the model is rotated by $\frac{2i\pi}{10}$ in frame f_i. In this way, memory cost are kept to a minimum and indeed, DNA strands of several millions base pairs easily fit in an average PC's RAM.

4.3 A hierarchy for interactive visualization and base-pair picking

The last ingredient of the data strucutre is a binary hierarchy over the sequence of frames $\{f_i, i \in [1, N]\}$. The root node consists of the entire sequence. The subsequence at one node is split in two sub-subsequences to form the two child nodes. At each node, a sphere bounding all the sampling points in the subsequence of the node is computed.

When drawing a long DNA strand, the hierarchy is traversed top-down, and if the bounding sphere of the node is sufficiently far away, we use alternative, simpler drawings of the sub sequence. Our implementation uses a double ribbon model for intermediate distances, and a simple line for faraway strands.

The hierarchy is also used to efficiently compute which base-pair has been clicked on by the user when he wants to manipulates the DNA, see below.

4.4 Twisting the DNA strand

Since our goal is the modeling of DNA strand in interaction with some proteins, it is important that the user user be able to fine tune the position of the DNA atoms; at least locally when there is a strong interaction with the outside world. Allowing independent atom placement, while most powerful, would destroy the curve model that we use to make DNA modeling easier. Instead, we allow the adjustment of the rotation of each base-pair around the curve. That is, at each frame f_i, the base pair is rotated by angle $\frac{2i\pi}{10} + a_i$, where a_i is specified by the user. More precisely, $a_i = 0$ initially, and the user can then modify the angle of each base-pair individually. The values are interpolated where they are not specifically defined: Suppose that the user has defined the values $\{a_{i_1}, a_{i_2}, ...a_{i_k}\}$ with $i_1 < i_2 < ... < i_k$. If $i_1 > 1$ we artificially set $i_0 = 1$ and $a_1 \leftarrow a_{i_1}$, and similarly after i_k. Then for all i, there exists j so that $i_j \leq i < i_{j+1}$, and the value a_i is a linear blend of a_{i_j} and $a_{i_{j+1}}$.

References

DAVEY, C. A., SARGENT, D. F., LUGER, K., MAEDER, A. W., , AND RICHMOND, T. J. 2002. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *Journal of Molecular Biology 319*, 5 (June), 1097–1113.

GOODSELL, D. S., AND JOHNSON, G. T. 2007. Filling in the gaps: artistic license in education and outreach. *PLoS Biology 5*, 12 (Dec.), e308.

JUNIER, I., MARTIN, O., AND KÉPÈS, F. 2010. Spatial and topological organization of DNA chains induced by gene co-localization. *PLoS Computational Biology 6*, 2 (Feb.), e1000678.

KUNKEL, T. A. 2004. DNA replication fidelity. *Journal of Biological Chemistry 279*, 17, 16895—-16898.

VENDEVILLE, A., LARIVIÈRE, D., AND FOURMENTIN, É. 2011. An inventory of the bacterial macromolecular components and their spatial organization. *FEMS Microbiology Reviews 35*, 2 (Mar.), 395–414.

WANG, W., JÜTTLER, B., ZHENG, D., AND LIU, Y. 2008. Computation of rotation minimizing frames. *ACM Transactions on Graphics 27*, 1 (Mar.), 2:1–18.

WINKLER, I., MARX, A. D., LARIVIÈRE, D., MANELYTE, L., GIRON-MONZON, L., HEINZE, R. J., CRISTOVAO, M., REUMER, A., CURTH, U., SIXMA, T. K., AND FRIEDHOFF, P. 2011. Chemical trapping of the dynamic MutS-MutL complex formed in DNA mismatch repair in escherichia coli. *Journal of Biological Chemistry 286*, 19 (May), 17326—17337.

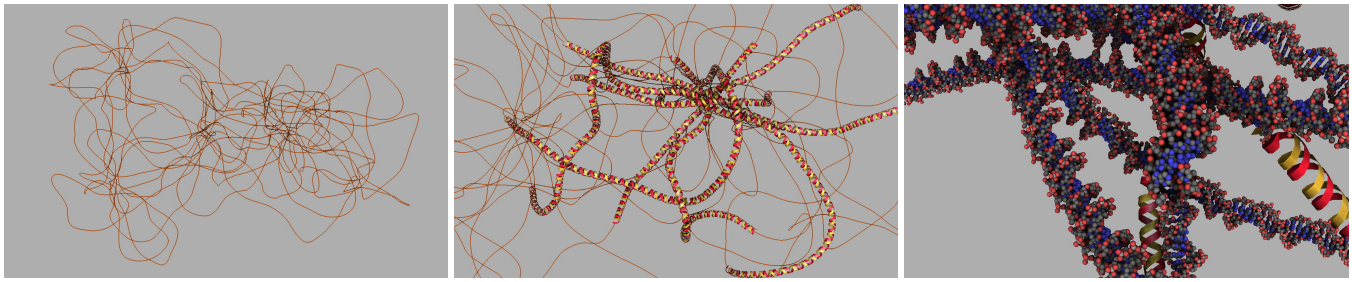

Figure 1: *Long DNA portion (28 000 DNA base pairs) visualized with levels of detail (see [Junier et al. 2010] for details about DNA compaction simulation).*

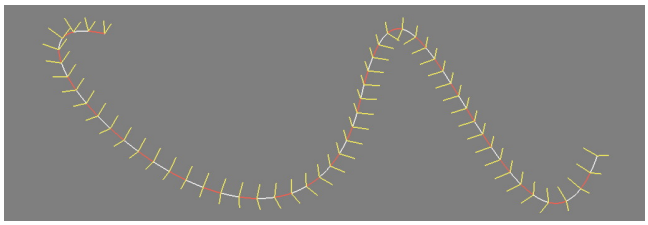

Figure 2: *A uniform, rotation minimizing frame sequence along a cubic Bézier curve. The normal and binormal vectors are shown yellow.*

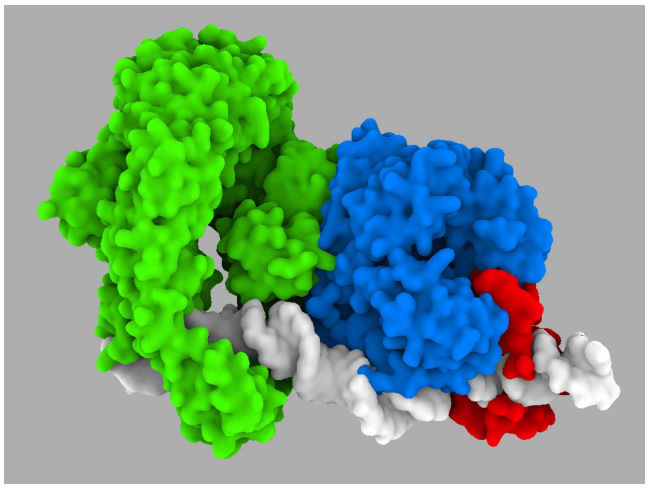

Figure 3: *A structural model for the arrangement of bacterial DNA repair proteins and DNA [Winkler et al. 2011].*

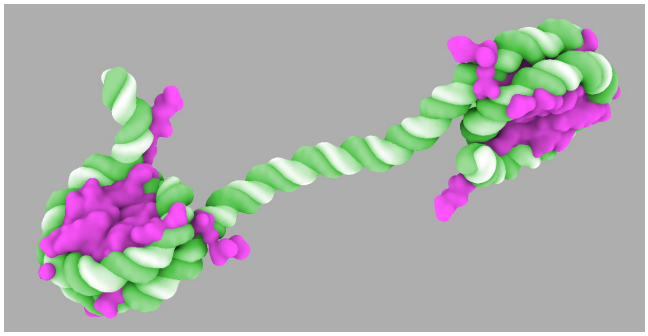

Figure 4: *A structural model for the wrapping of DNA around two nucleosomal proteins [Davey et al. 2002].*

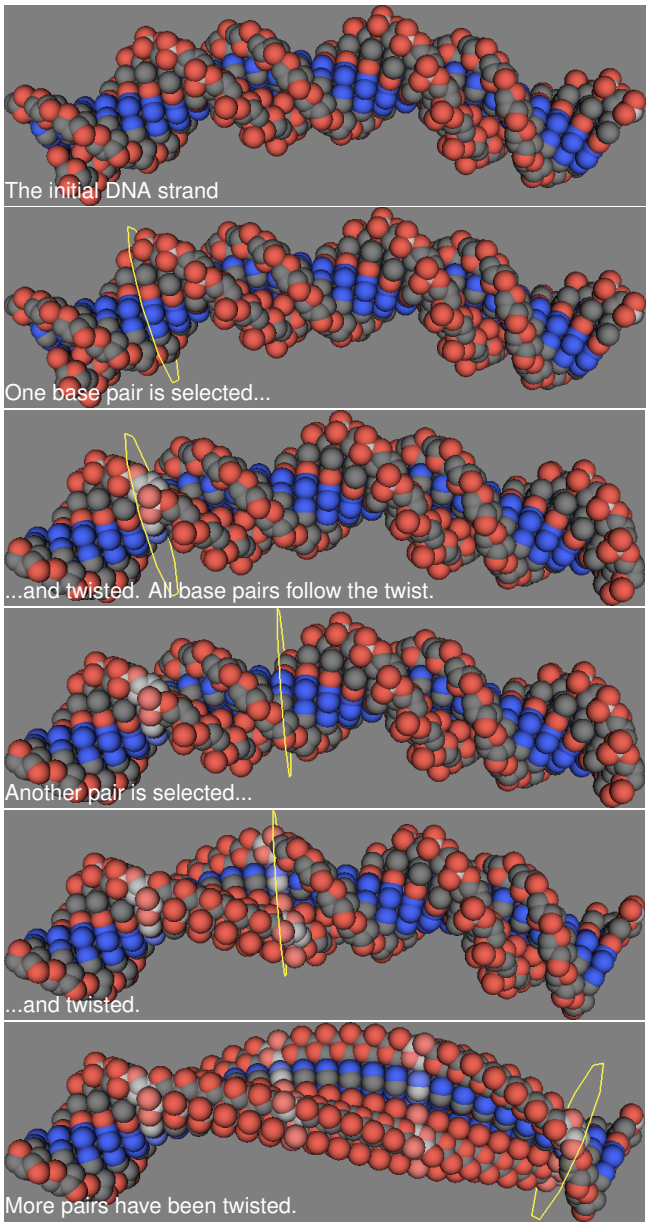

Figure 5: *Twisting DNA.*