Pacific Symposium on Biocomputing 7:151-162 (2002)

REPRESENTATION AND PROCESSING OF COMPLEX DNA SPATIAL ARCHITECTURE AND ITS ANNOTATED GENOMIC CONTENT

RACHID GHERBI AND JOAN HERISSON

Gesture and Image group
LIMSI-CNRS, Université Paris-Sud,BP 133, F-91403 ORSAY CEDEX, FRANCE*
http://www.limsi.fr
E-mail: gherbi@limsi.fr, herisson@limsi.fr
Phone: +33 1 69 85 81 64 or 66
Fax: +33 1 69 85 80 88

This paper presents a new general approach for the spatial representation and visualization of DNA molecule and its annotated information. This approach is based on a biological 3D model that predicts the complex spatial trajectory of huge naked DNA. With such modeling, a global vision of the sequence is possible, which is different and complementary to other representations as textual, linguistics or syntactic ones. The DNA is well known as a three-dimensional structure. Whereas, the spatial information plays a great part during its evolution and its interaction with the other biological elements This work will motivate investigations in order to launch new bioinformatics studies for the analysis of the spatial architecture of the genome. Besides, in order to obtain a friendly interactive visualization, a powerful graphic modeling is proposed including DNA complex trajectory management and its annotated-based content structuring. The paper describes spatial architecture modeling, with consideration of both biological and computational constraints. This work is implemented through a powerful graphic software tool, named ADN-Viewer. Several examples of visualization are shown for various organisms and biological elements.

1 Introduction and state of the art

People use very often the standard textual format (or a near representation), considering DNA as a succession of letters (A, C, G or T) that represent the nucleotides of the molecule. With such plat representation, it is very hard to perceive any global pertinent information of DNA sequence. Even if t is possible to structure the sequence as a hypertext [3.], the user has finally only a local textual point of view at each level of the document.

The DNA is well known as a three-dimensional structure [10.]. Whereas, the spatial information plays a great part during its evolution and its interaction with the other biological elements [4. ,5. ,8. ,11. ,12. ,13. ,14. ,15. ,16. ,17. ,18. ,19.].

In this context, it appears essential to design software tools focused on the representation, visualization and interactive exploration of the three-dimensional information of DNA. Besides, it is necessary to add functionalities in order to perform quantitative measures and to systemize the processing of various types of spatial information (curvature, compactness, geometric distances, etc.). Based on biological 3D conformation models of naked DNA, this paper presents our work aiming at building virtual three-dimensional information from DNA sequences. This

construction will allow biologists to visually scan and characterize the spatial architecture of DNA of any size. This involves computational consideration, in particular for computer graphics algorithms (scene management algorithms, user interaction facilities, reliable and pertinent visualization and representation, etc.).
Some present and past Biological studies are concerned with the 3D structure of DNA. However, these studies are very specific to particular sequence elements and limited to small size of DNA sequences. The most used model is Bolshoy and Trifonov one [4.]. From computer science point of view, a few tools can visualize DNA sequences. They operate using a prediction algorithm, but they can visualize sequences that do not exceed on thousand of nucleotides (700 for *DNAtools*© [7.]). Many of these tools are developed by biologists teams and are exclusively dedicated to some particular biological problems. As far we know, there is no application that allows representing the spatial conformation of complete chromosomes.

This paper firstly explains the mechanism of the trajectory prediction. Then, a description about graphical representation and visualization of DNA sequences is presented. Some of the user interface functionalities are listed and finally many examples of visualization for different organisms are shown.

2 Biological interest of the DNA spatial information

The analysis of the nucleic sequences is a problem of a great complexity because of the superposition of various "signals" in the DNA. It is necessary to carry out genomic analysis with various and complementary approaches. The interest to model the structural aspects of the nucleic acids was charged for a long time and significant progress were made during these last years. This subject, far from becoming exhausted, does not stop growing rich starting from the experimental knowledge acquired on the DNA [11. ,4. ,8. ,13.].

One cannot define in an exact and complete way the existence of a functional promoter. Thus, even for simple models as *S. cerevisiae*, it is difficult to affirm by data processing analysis that an ORF (Open Reading Frame) of small size is actually expressed. Several assumptions explaining the putative role of the curve in the regulation regions exist now. The curved DNA can form large loops around the RNA polymerase, and thus increases the affinity of the DNA complex. It was shown that even a very low curve could increase enormously the affinity of the connection proteins-DNA, which led Suzuki and Yagi [15.] to put forth the assumption that the local curve can be used to precisely adjust the forces of interaction between the promoters and the regulation factors. It was also supposed that the curve of the DNA gathered the components of the transcriptional complex, spread out along the molecule of DNA. The curve and/or the structure in super-helix of the DNA results from a torsion stress which affects the energy of fusion of the DNA and the unfolding of the double helix, thus making it possible to assist (even to replace) corresponding initiating proteins.

3 Prediction of the 3D trajectory of DNA

DNA is a double strands (*Watson* and *Crick* [10.]) structure representing a double helix. The strands are anti-parallel and complementary (A is associated with T and C with G and *vice versa*. This association is called *base pair* (*bp*).

To build the 3D trajectory of DNA, two kinds of input data are necessary: DNA sequence in textual format and a 3D conformation model [1.]. Several 3D models exist. The most used are Bolshoy's [4.] and Cacchione's [9.] ones. The algorithm used for the 3D prediction of the trajectory of DNA is described in a previous work [1.].

4 Graphical representation and user interaction

4.1 Genomic and genic representations

When the sequence length is greater than few hundreds nucleotides or the observer is far from the DNA, he does not need to see all its details, but only some global information (shape, compactness, its curvature, etc.). In addition, for huge sequences (cf. figures pages 7, 8, 9 and 10), it is essential to reduce to the utmost of one's ability the representation. That is why, in genomic representation, we represent a nucleotides by a simple successive linked points and only the two strands are displayed (cf. Figure 1).

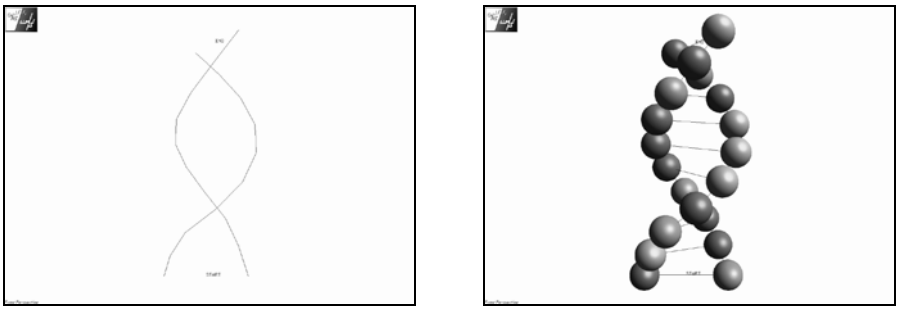

Figure 1. Sequence of 11 bp: genomic representation on the left, genic one on the right.

If the user wants to visualize or access to a little part of the sequence, he needs to see more details than in genomic representation. The genic representation provides supplementary information on nucleotides themselves. Indeed, each nucleotide is represented by a colored sphere (cf. Figure 1). This representation is suitable to visualize DNA sequences as genes, transposons, promoters regions, etc.

4.2 User interaction

ADN-Viewer integrates several powerful and friendly interaction capabilities. The user can manipulate and animate (geometric rotations and translations) the molecule using the mouse. Besides, the keyboard allows the user to control many parameters within the application. Some of interaction functionalities are for example fragment extraction (cf. Figure 2), the choice of any of existing conformation models, etc.

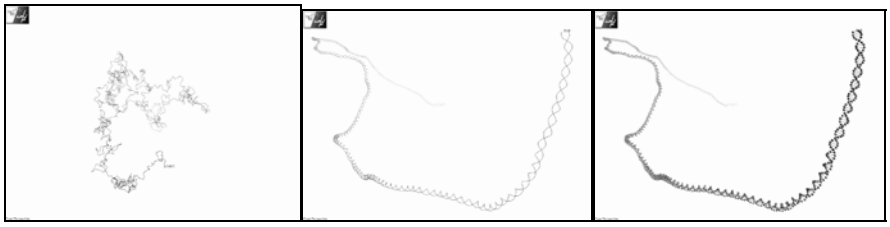

Figure 2. Sequence of 300 Kbp and an extracted part in genomic & genic representations.

5 Visualization and scene management

Several modes of visualization are provided within *ADN-Viewer*. The user can visualize a sequence according to different points of view. This functionality is very useful because, in general way, two 3D objects can have the same spatial properties in one view whereas they are different in the other ones. *ADN-Viewer* offers four different points of view: a front view using perspective projection; a front, left side and under views using orthogonal projection.

One can also visualize the molecule bulk in the three-dimensional space by displaying its bounding box (cf. Figure 3). This functionality is also used to compute quantitative compactness values (span, volume, density, etc.).

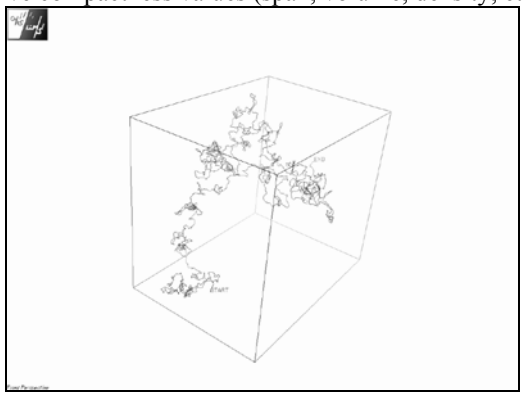

Figure 3. The bounding box gives bulk information on the sequence.

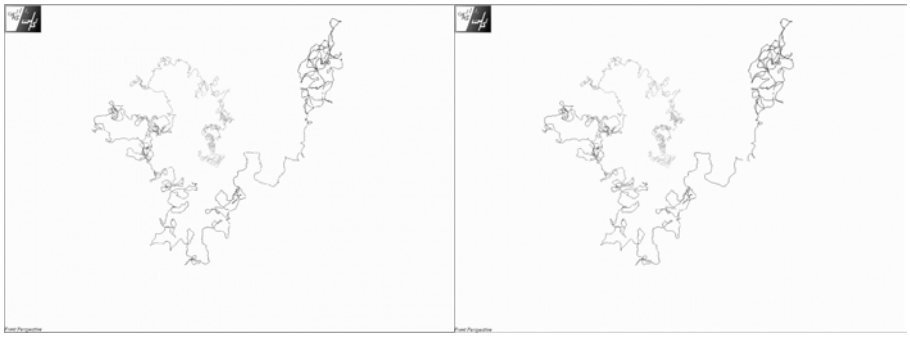

Figure 4. The *Yeast chrIII* original sequence and its sampled one by a factor value 20.

If the 16th chromosome of *Yeast* is considered as an example, it counts about 948061 bp. This implies to display about two millions of nucleotides (2 nucleotides per plate) at each movement of the molecule performed by the user. As we search a real time animation and interaction with the molecule, it is unthinkable to visualize such amount of points on classic workstation. It is thus necessary to filter as possible as the graphical information flow. Nevertheless, this process must not modify the trajectory of DNA. For the two kinds of representations (cf. section 4), sampling algorithms are applied on the sequence.

In the case of genomic representation, only the global shape of the trajectory is taken into account. The goal of the sampling algorithm is to reduce the number of displayed. The difficulty is to find the best sampling value that does not modify the trajectory but reduces the number of points allowing ADN-Viewer to function in interactive time. Presently, if the user changes the point of view of the DNA we must adapt the sampling according to the distance that separates the observer from the molecule. The Figure 4 illustrates an example of considerable sampling that does not modify the three-dimensional structure of the DNA.

For the genic representation, we use the same previous algorithm but applied on the number of displayed spheres and on their level of detail. This is also done according to the observer-molecule distance.

6 Modeling and visualization of annotated DNA sequences

The visualization of DNA sequence is very fruitful but not enough and it is necessary to carry out quantitative studies on the DNA molecule and its contents (density of genes, compactness cartography, curvature maps, spatial distribution, DNA-proteins interaction...). The trajectory of the DNA can inform us about its contents (genes, introns, exons, transposons...). Within the framework of this paper, we are interested in particular on genes in order to investigate their spatial

relationship. The proteins that come to be fixed on a gene (upstream of it) for the transcription phase need to reach it directly. They must thus pass through the "swell of wool" formed by the DNA molecule. It thus appeared to us interesting to study the space of manœuvre available to these proteins to reach their target. For that, we are undertaking a comparative study of the distances that separate genes in order to reveal which areas are rich in genes. We can also study the correlation between parametric distances and spatial ones along the sequence. This work is in progress and some first results are shown by Figure 6 for genes content within chromosomes.

The visualization of DNA content needs interfacing ADN-Viewer with annotated accessible genomic databases or databanks. Nevertheless, there is no standard format used to describe a genomic sequence. In order to temporary bypass this problem, ADN-Viewer was interfaced with an ad-hoc database that contains only annotations of genes of some organisms. This solution presents two advantages compared to the online banks: the reliability of the data enabling ADN-Viewer to receive sure formatted information from the base, and its opening potentialities towards any existing or future structured databases. ADN-Viewer allows the user to access all the annotated information of the base through an explorer and to visualize them. When a sequence is selected, a level is added in the tree structure of the explorer and the user can access to all corresponding genes.

In the other side, ADN-Viewer will be able to augment the base by providing it pictures of the molecules as well as geometrical and spatial information.

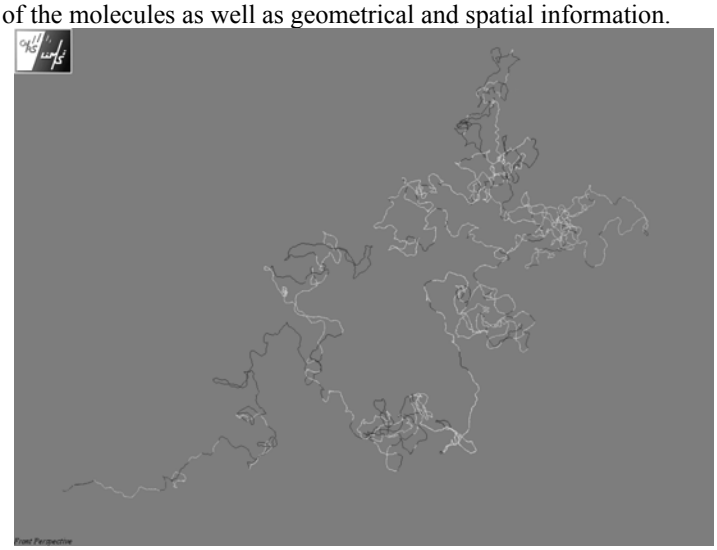

Figure 5. Visualization of genes content for *S. cerevisiae chrI*. Genes regions are white colored and intergenic regions are black colored.

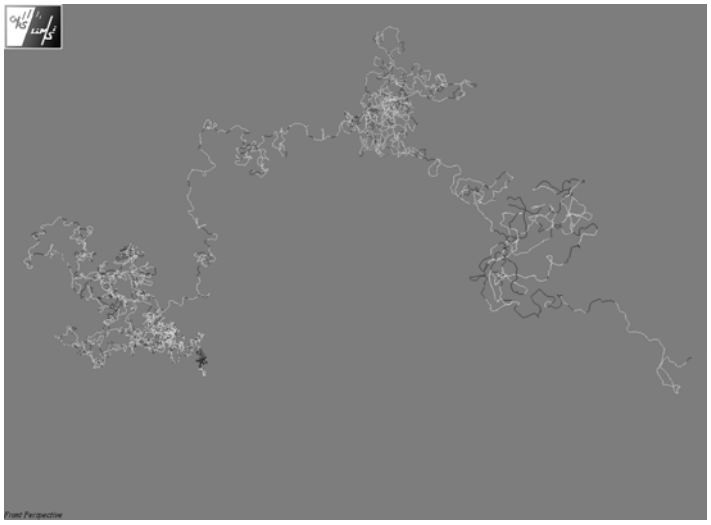

Figure 6. Visualization of genes content for *S. cerevisiae chrIV* (bottom view). Genes regions are white colored and intergenic regions are black colored.

7 Examples of visualization of genomes and biologic elements

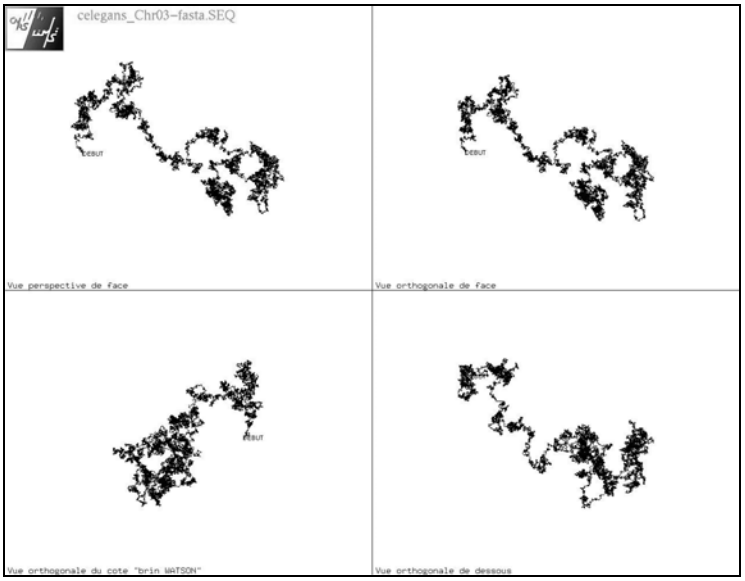

Figure 7. C. elegans_chr3 (11850213 bp).

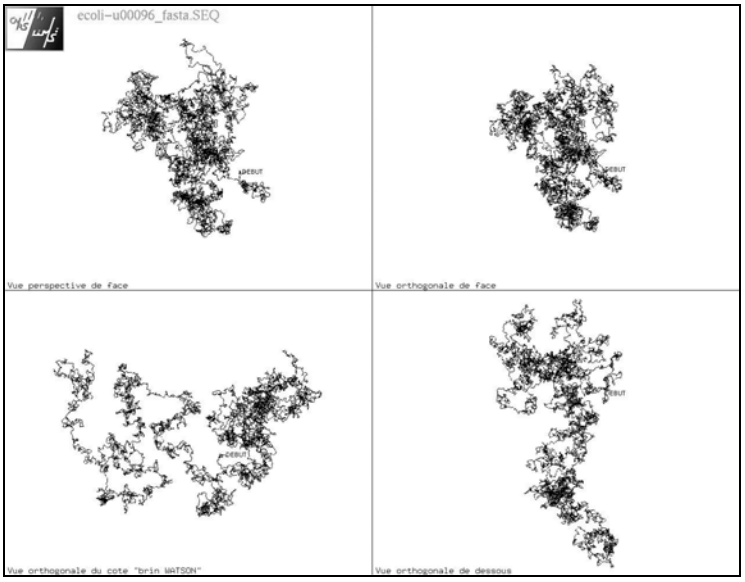

Figure 8. E. coli (4639221 bp).

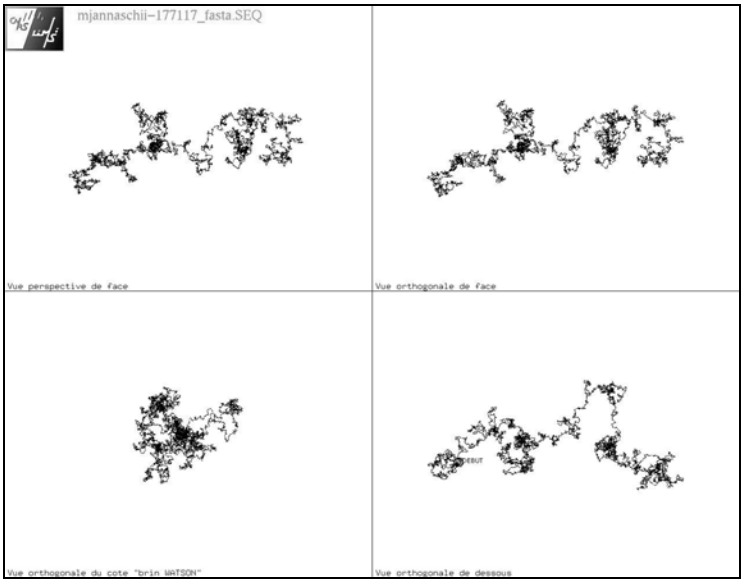

Figure 9. M. jannaschi (1664957 bp).

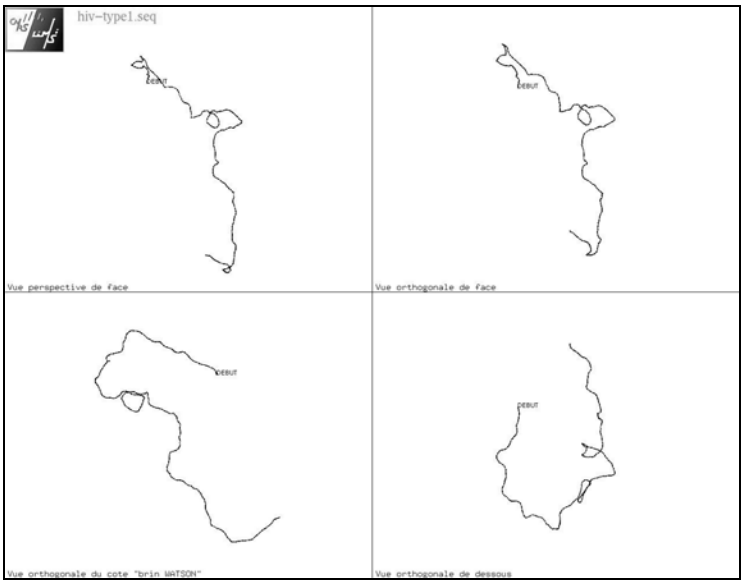

Figure 10. HIV_type1 (9181 bp).

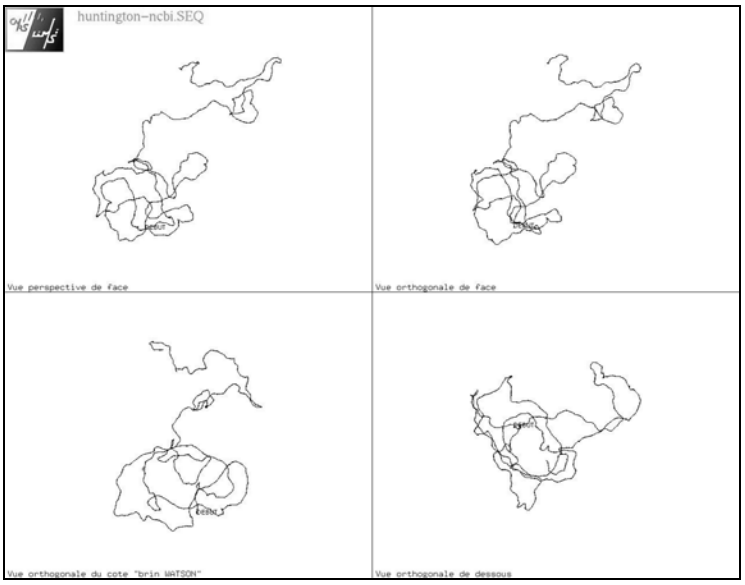

Figure 11. Huntington (55204 bp).

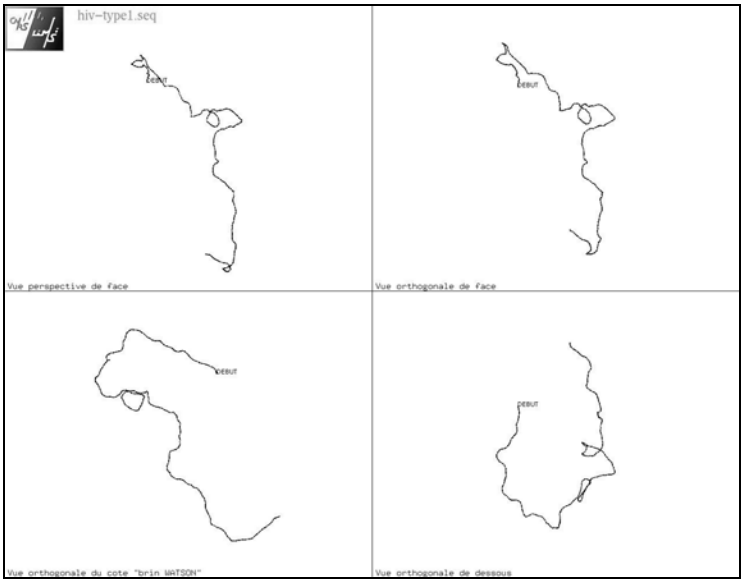

Figure 12. Fot1 (1928 bp).

8 Conclusion and future work

The 3D representation and visualization of DNA molecule and interacting with it make it possible to have a global point of view of the sequence, in opposition with the textual format. This brings a new vision and an original approach suitable to launch new bioinformatics studies for the analysis of the genome. Besides, in order to obtain a friendly powerful interactive visualization, a dedicated modeling process is essential and various representations are necessary in order to adapt the visualization according to different analysis cases. In this paper, we carried out work about the spatial architecture of naked DNA, with consideration of both biological and computational constraints. A graphic software tool, named ADN-Viewer, implements this work.

Besides, it was interesting to represent and visualize the annotated biological elements of DNA sequences. To analyze spatial relations between genomic elements, it was possible to display the genes of chromosomes. Nevertheless, this implied to interface *ADN-Viewer* with genomic database and to structure the graphical representations for a better visualization and real-time user interfacing More generally, one can have a global vision of all annotated information in order to study various relations of genes, promoters, terminators, transposons, etc.

Presently and in a future work, various algorithms are in design, dealing with curvature maps computation, geometric distances estimation, compactness cartography elaboration, etc.

We are working now to build a software platform, called **GenoMEDIA**, that integrates the graphical and interactive tool *ADN-Viewer* in its server implementation, a genomic database, and a web server that dialog with the previous both tools. In the close future, a first version of **GenoMEDIA** will be open on the Internet and can be accessible by any user using his own navigator.

For the long-term research, the first work that will be carried out is the validation and comparison of the various conformation models in order to obtain the most reliable prediction. This implies to do different matching and fitting processes between the predicted images and real ones that could be obtained with advanced electronic microscopy.

Acknowledgements

We thank the following academic institutions and people for their financial and advising help: LIMSI-CNRS, Paris-Sud University, the French bioinformatics program inter-EPST CNRS-INRIA-INRA-INSERM, and People from IGM and IBP form Orsay, Monique Marilley from Marseille, Ed. Trifonov from Rehovot university in Israel.

References

1. Rachid Gherbi and Joan Hérisson. ADN_Viewer, a software framework for 3d modeling and stereoscopic visualization of the genome. In Proc. of Graphicon'2000, International conference on Computer Graphics and Vision, Moscow, August-September 2000.
2. Henn C. and Teschner, "Interactive molecular Visualization", Track session at PSB ???
3. Grigoriev A., "Reusable graphical interface to genome information resources", in proc. of PSB conference ???
4. A. Bolshoy, P. McNamara, R.E. Harrington, and E.N. Trifonov. Curved DNA without A-A: Experimental estimation of all 16 DNA wedge angles. *Proc. Natl. Acad. Sci. USA*, 88: 2312-2316, March 1991.
5. Philippe Pasero. *L'organisation du chromosome eucaryote et ses implications dans le contrôle de l'activité génique et la transmission des patrons d'expression*. PhD thesis, University of Aix-Marseille II, Faculty of Sciences of Luminy, December 1993.

6. Marilley M, Pasero P (1996). *Nucleic Acids Res.* **35**:2204-2211
7. Søren Wilken Rasmusen. DNAtools©. http://www.dnatools.org.
8. P. De Santis, A. Palleschi, M. Savino, and A. Scipioni. A theorical model of DNA curvature. *Biophys. Chem.*, 32: 305-317, 1988.
9. Cacchione S, De Santis P, Foti DP, Palleschi A, Savino M (1989). *Biochemistry* **28**, 8706-8713
10. J.D. Watson and F.H.C. Crick. Molecular structure of nucleic acids. *Nature*, 171: 737-738, April 1953.
11. I. Lafontaine and R. Lavery, Curr. Opin. Struct Biol.1999, 9:170-176.
12. Matthews KS (1992). *Microbiology Reviews* **56**:123-136
13. Ponomarenko M.P., Ponomarenko J.V., Kel A.E. and Kolchanov N.A., "Search for DNA conformational features for functional sites. Investigation of the TATA box", in proc. of PSB conference, ?????
14. Perez-Martin J, Rojo F, de Lorenzo V (1994). *Microbiology Reviews* **58**:268-290
15. Suzuki M, Yagi N (1995). *Nucleic Acids Res.* **23**:2083-2091
16. Natale DA, Umek RM, Kowalski D (1993). *Nucleic Acids Res.* **21**:551-560
17. Nickerson CA, Achberger EC (1995). *Journal of Bacteriology* **177**(20): 5756-5761
18. Carmona M, Claverie-Martin F, Magasanik B (1997). *Proc Natl Acad Sci USA* **94**:9568-9572
19. Beloin C, Exley R, Mahé AL, Zouine M, Cubasch S, Le Hégarat F (2000). *Journal of Bacteriology* **182**(16): 4414-4424

---

* LIMSI-CNRS is a laboratory of informatics for mechanics and engineering science, from the French National Centre of Scientific Research, located in a south of Paris in the campus of Paris-South University.