Parallelizable Strategy for the Estimation of the 3D Structure of Biological Macromolecules

Claudia Caudai*, Monica Zoppé†, Emanuele Salerno*, Ivan Merelli‡ and Anna Tonazzini*
*CNR, Institute of Information Science and Technologies, Via Moruzzi, 1, 56124 Pisa, Italy
Email: claudia.caudai@isti.cnr.it
†CNR, Institute of Clinical Physiology, Via Moruzzi, 1, 56124 Pisa, Italy
‡CNR, Institute of Biomedical Technologies, Via Fratelli Cervi, 93, 20090 Segrate (MI), Italy

***Abstract*—We present a parallelizzable, multilevel algorithm for the study of three-dimensional structure of biological macromolecules, applied to two fundamental topics: the 3D reconstruction of Chromatin and the elaboration of motion of proteins. For Chromatin, starting from contact data obtained through Chromosome Conformation Capture techniques, our method first subdivides the data matrix in biologically relevant blocks, and then treats them separately, at several levels, depending on the initial data resolution. The result is a family of configurations for the entire fiber, each one compatible with both experimental data and prior knowledge about specific genomes. For Proteins, the method is conceived as a solution for the problem of identifying motion and alternative conformations to the deposited structures. The algorithm, using quaternions, processes the main chain and the aminoacid side chains independently; it then exploits a Monte Carlo method for selection of biologically acceptable conformations, based on energy evaluation, and finally returns a family of conformations and of trajectories at single atom resolution.**

***Keywords*-Parallel Computing; Macromolecules; Hierarchical Reconstruction; Chromatin; Protein Motion.**

I. Introduction

The need to observe the structure of biological macromolecules at multiple resolution levels emerged in recent years, following major advances in investigating the three-dimensional structure of proteins and nucleic acids. A primary example of this approach is represented by the study of chromatin, which resulted in the identification of TADs (Topologically Associating Domains) [1], portions of chromatin with many internal and few external interactions, described as fairly isolated structures within the fibre. For the 3D reconstruction of chromatin structure using Chromosome Conformation Capture and other HI-C data [2]–[4] it is therefore possible and convenient to deal in parallel with the TAD structures and then reconstruct and investigate the structure of the chromatin in its entirety, adopting a coarse-grained approach that follows the most natural division.

Another important topic in biology is the recognition of the dynamic nature of protein activity. Because at this time it is impossible to observe protein movements experimentally, it is relevant to speed up the methods for the analysis of their structural and conformational changes. Treating in parallel autonomous or partially independent sub-structures is a smart strategy for the analysis of three-dimensional biological structures. For example, in the case of proteins, the subdivision of the chain in amino acids provides the smallest units for computation.

This approach leads to consider biological phenomena at multiple levels of resolution. Organic structures are often fractal, which is why coarse-grained methods are successful in biology. Even better than coarse-grained methods, multilevel methods can be used to observe biological structures at different resolutions. The organization of chromatin into TADs, for example, leads to overcome the two-dimensional model, and to the description of DNA as a pattern of nested structures in which TADs and sub-TADs form a complex hierarchy. [5].

Following this approach, we propose an iterative and multilevel parallelizable algorithm for the investigation of the three-dimensional structure of macromolecules.

II. Method

The method we propose, whose code flow is illustrated in Algorithm 1, is a parallelizable and multilevel algorithm. The molecular chain is modelled as a bead-chain, maintaining the biological order of the sub-structures. The algorithm always starts from the highest possible resolution, named INPUTCHAIN in Algorithm 1, and subdivides the global structure into sub-structures trying to respect as much as possible the natural subdivisions already present. Sub-structures are processed in parallel, looking for configurations compatible with a-priori constraints and with low potential energy. The molecular chain is then recomposed and evaluated from a structural and energetic point of view. If successive levels of resolution exist, the chain is modelled with coarse-grained techniques and the algorithm repeats, as illustrated in Algorithm 1. At the end of the reconstruction a final configuration is proposed, named OUTPUTCHAIN in Algorithm 1. The coordinates of all the sub-structures are recovered, up to the highest resolution, so as to provide a tool to explore macromolecules in their entirety at multiple levels of resolution.

Algorithm 1 Code Flow

Input: inputchain

The input is the molecular chain at the highest resolution

nlevel= 0

Initialisation at level=0:

inputchain=C

nchains=[len(C)]

while nchains[nlevel]>1 **do**

  subchains=dividechain(nlevel,C)

  C *is a string and* subchains *is an array of strings*

  nchains[nlevel]=len(subchains)

  nchains.append(nchains[nlevel])

  Parallel Part:

  for i in range(nchains[nlevel]) **do**

    C_i =annealing(subchains[i])

  end for

  Increase level:

  nlevel=nlevel+1

end while

Compute coordinates of the highest resolution chain:

outputchain=compose(C_i)

return outputchain

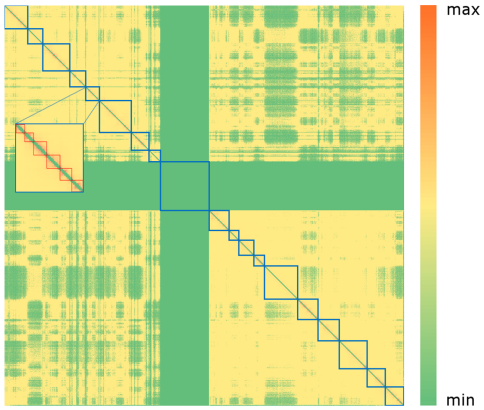

Figure 1. Heatmap of contact frequency matrix of Chromosome 16 at 100kbp resolution. Multilevel subdivision in diagonal blocks was performed by our algorithm. Note the repeating pattern in the zoom of a block.

III. Experiments

Our method is suited for both the 3D reconstruction of chromatin, starting from Hi-C data, and for the elaboration of Molecular Dynamics of proteins [6], as shown below.

A. 3D Chromatin Conformation

To obtain 3D Chromatin conformations consistent with information stored in contact frequency matrices derived from Hi-C experiments [7], we introduce a chromatin model consisting of a chain of consecutive and partially penetrable beads whose properties (bead size, elasticity, curvature, etc.) are suitably constrained. Our algorithm [8] automatically divides the contact matrix in diagonal blocks, dividing the chain into sub-chains corresponding to TADs and sub-TADs. The method is parallel and iterative: starting from the highest resolution level, the algorithm elaborates the smallest sub-chains in parallel, using a Monte Carlo method [9]; it then re-iterates on the chain at successive resolution, in which conformations obtained in the previous level are modeled as beads with coarse-grain methods. Perturbations and movements of the chains are perfomed by using quaternions, a mathematical device more efficient than Euler matrices in managing rotations, as fully explained in [10]. In the algorithm the data-fit funcion rewards the proximity of beads with high contact frequencies and avoids deep interpenetartions. The process is repeated for every resolution level up to the lowest one. At the end the whole chain is modelled and the structure can be investigated at the highest possible resolution. Starting from a contact matrix, several final conformations are produced, in order to sample the space of possible conformations fitting with initial contact information. A fundamental step in the analysis and processing of data from Hi-C experiments is the division in blocks of the contact matrix. This subdivision is based on an intrinsic property of DNA, which can be conceptually subdivided into contiguous domains, called TADs, consisting of compact spatial modules. This pattern suggests that the spatial organization of DNA follows a hierarchical model. In Figure 1 a heat-map of the contact matrix at 100kbp resolution of chormosome 16, obtained from freely available Hi-C data of Rao [11] is represented; diagonal blocks with high contact frequency can be easily distinguished by eye; a zoom of one of these blocks shows that this pattern is multilevel. The TAD identification step is crucial in 3D Chromatin reconstruction and can be performed following different principles; our block detection method is based on the algorithm of [12], and leads to a nested series of blocks, reflecting the hierarchical nature of Chromatin compaction in cellular nuclei [13]. As an example, in Figure 2 we report some final configurations of Chromosome 16, obtained starting form contact matrices binned at different resolutions. The higher the resolution, the higher the size of the contact matrix and consequently the compuational complexity of the system. In the most complex scenario, the contact matrix can be divided into numerous small blocks, and at several levels. These are computed in parallel,with each sub-chain computed independently, in order to reduce calculation times, as illustrated in Table I. By computing every sub-chain in parallel with different processors, calculation times decrease fast, because on one side there is no exchange of information within the sub-chains of each level, and on the other side most of the CPU time is used for sub-chains calculations. Treating sub-chains in series the total elapsed time is the sum of elapsed times of every sub-chain of every resolution level. By contrast, computing sub-chains in parallel (in different processors),

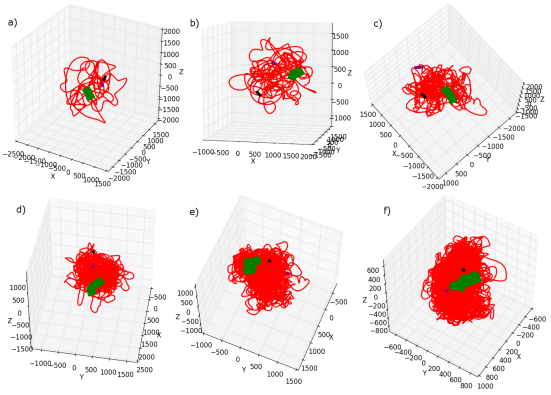

Figure 2. Examples of final configuration of Chromosome 16 obtained from [11] at different resolutions: a) 500kbp, b) 250kbp, c) 100kbp, d) 50kbp, e) 25kbp and f) 10kbp resolution. In every panel, the centromere is highlighted in green. Scales are in nm.

Table I
CALCULATION TIMES IN SERIES AND IN PARALLEL

		elapsed time[b] (in secs)	
	num of sub-chains [a]	in series	in parallel
Chromosome 16			
resolution 500kbp	19	4502	1030
resolution 250kbp	33	11389	1050
resolution 100kbp	91	19288	1420
resolution 50kbp	187	35903	3670
resolution 25kbp	367	70092	3980
resolution 10kbp	961	160703	3540
1CFC	149	14825	4960

[a] sum of sub-chains in all resolution levels.
[b] processor 2X Intel(R)Xeon(R) 40 core,16Mb GPU,128Gb RAM.

the total elapsed time will be the sum of maxima of all elapsed times for every resolution level.

B. Molecular Dynamics

Our method can also be applied to the investigation of molecular dynamics (MD) of proteins. Classical MD simulations provide detailed information on the conformational changes of proteins [6]. Positions and movements are found as very expensive solutions to complicated differential equations. Our approach to MD combines quaternions, to manage the movements of atoms on their own trajectories, and Monte Carlo Methods [9], to perform incremental rotations and control energy values. In our model the protein is treated both at the level of main chain, as a chain of successive beads, N-CA-C, and at the level of side chains, as short chains of successive beads, coresponding to heavy atoms. At each step the backbone and every amino acid are perturbed in parallel. Their interactions are managed

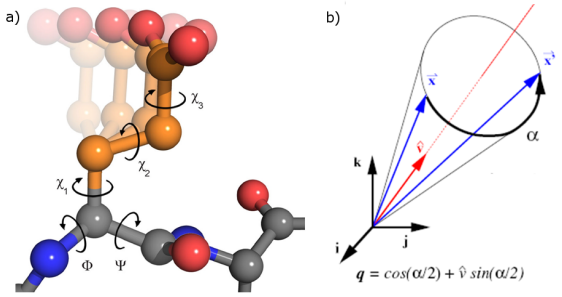

Figure 3. a) Three-dimensional representation of Glutamic Acid. χ_i are the side chain dihedral angles, Φ and Ψ are the back-bone angles. Figure from [15]. b) Graphical representation of the spatial transformation of the vector $\vec{x}$ in the vector $\vec{x}'$ through the composition with the quaternion q. By composition with the quaternion q, the vector $\vec{x}$ performs a rotation of α around the axis whose direction is defined by the versor $\hat{v}$.

evaluating energetic field [14]. In this way, trajectories of proteins can be easily evaluated from an energetic point of view, thus offering a faster alternative to classical MD. From a geometrical perspective, proteins are treated as kinematic chains totally defined by the position of atoms (3D coordinates) and segments with fixed length joining consecutive atoms (peptide and chemical bonds), as shown in Figure 3.

In our model we exploit the fact that atom trajectories are regular and well defined: every atom can rotate along a precise cicular tajectory, defined by Lennard Jones forces, around the axis identified by the two previous atoms. The algorithm manages the evolution of the sistem by Simulated Annealing, using quaternion algebra for random incremental rotations, perturbing atom positions and leaving bond-lenghts fixed. The algorithm treats every subchain indipendently, but the energetic evaluation is performed over the whole chain. For energy calculation we use the AMBER99 force-field database [14]. Starting from the same PDB file, different runs produce different configurations, in order to explore other possible conformations with low potential energy, for the protein under study. Once a new conformation is found, the trajectory from the starting to the final conformation can be observed.

For our experiments we used Calmodulin, a very flexible protein composed of 148 amino acids (1165 heavy atoms). The PDB file 1CFC [16] contains 25 stable conformations derived from NMR experiment. In Figure 4a, the final conformation produced by one run of our program, starting from stable conformation 1, is shown. The output conformation, with an RMSD of 4.8 Å from the starting conformation, is very close to conformation 10 of PDB file (2,4 Å of RMSD). In 100 runs, 100 different final conformations are obtained; these final configurations have been aligned with all the 25 stable configurations of 1CFC and RMSD have been calculated. Figure 4b sums up results: starting from stable conformation 1, output conformations of our program

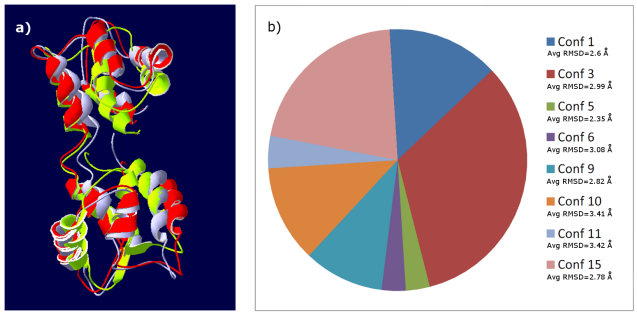

Figure 4. a) Structures of Calmodulin: in yellow the conformation 1 of the PDB file 1CFC, used as input of our program; in red one of the output configurations; in light violet the conformation 10 of the PDB file 1CFC, very similar to the output configuration. b) Pie-plot of the correspondence of the outputs of 100 runs with stable conformations of Calmodulin found in PDB file 1CFC.

result close (RMSD between 2.6 and 3.4 Å) to 8 stable conformations present in the 1CFC PDB file.

Treating every aminoacid and back-bone in parallel with different processors, calculation times decrease, as shown in Table I. Reducing calculation times in MD is challenging and we propose a smart and fast approach to the investigation of the possible trajectories and conformation changes of proteins.

IV. Conclusion

In conclusion, we present here an algorithmic method to compute some of the most challenging and interesting aspects of biology: chromatin structure and protein dynamics. The methods, despite the obvious differences related to the nature of peptides and nucleic acids, have some features in common, including the use of quaternions to introduce rotations, and, most prominent, the possibility of computing sub-structures in an independent, and hence parallelizable, way. For DNA the smallest units are the smallest TADs that can be identified from the contact matrix; these can be so small, that a second, third and higher order of TAD might be necessary for the 3D reconstruction of the complete sequence.

For proteins the smallest units are the single aminoacids, each of which is characterized by its specific set of parameters, derived from the study of known protein structures. The massive parallelization introduced by the method, allows for a significant reduction in the computation time for each process under study, making it possible to produce a large number of solutions, as it is common (and necessary) in biological studies. At the same time, splitting the problems into smaller units does not preclude the incorporation of further information at the successive levels, as demonstrated in the study of human Chromosome 16.

Acknowledgment

This work has been partially supported by the Italian Flagship Project InterOmics, WP01-ISTI, and by ISTI-CNR, through scientific agreement with ITB-CNR, Milan.

References

[1] J. R. Dixon *et al.*, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, pp. 376–380, 2012.

[2] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, "Capturing chromosome conformation," *Science*, vol. 295, pp. 1306–1311, 2002.

[3] J. Fraser *et al.*, "Chromatin conformation signatures of cellular differentiation," *Genome Biology*, vol. 10, p. 37, 2009.

[4] S. Wang, J. Xu, and J. Zeng, "Inferential modeling of 3d chromatin structure," *Nucl. Ac. Res.*, vol. 43, p. e54, 2015.

[5] M. Forcato, C. Nicoletti, P. Koustav, C. M. Livi, F. Ferrari, and S. Bicciato, "Comparison of computational methods for hi-c data analysis," *Nature Methods*, vol. 14, p. 679685, 2017.

[6] J. D. Hirst, D. R. Glowacki, and M. Baaden, "Molecular simulations and visualization: introduction and overview," *Faraday Discuss*, vol. 169, pp. 9–22, 2014.

[7] E. Lieberman-Aiden *et al.*, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science*, vol. 326, pp. 289–293, 2009.

[8] C. Caudai, E. Salerno, M. Zoppè, and A. Tonazzini, "Inferring 3d chromatin structure using a multiscale approach based on quaternions," *BMC Bioinformatics*, vol. 16, p. 234, 2015.

[9] M. Rousseau, J. Fraser, M. A. Ferraiuolo, J. Dostie, and M. Blanchette, "Three-dimensional modeling of chromatin structure from interaction frequency data using markov chain monte carlo sampling," *BMC Bioinf.*, vol. 12, pp. 414–429, 2011.

[10] C. F. Karney, "Quaternions in molecular modeling," *J. Mol. Graph. Model.*, vol. 25, pp. 595–604, 2007.

[11] S. S. Rao *et al.*, *Cell*, vol. 159, p. 16651680, 2014.

[12] B. R. Lajoie, J. Dekker, and N. Kaplan, "The hitchhiker's guide ti hi-c analysis: Practical guidelines," *Methods*, vol. 72, pp. 65–75, 2015.

[13] C. Caudai, E. Salerno, M. Zoppè, and A. Tonazzini, "Estimation of the spatial chromatin structure based on a multiresolution bead-chain model," *IEEE/ACM TBCC*, 2018, in press, doi: 10.1109/TCBB.2018.2791439.

[14] W. D. Cornell *et al.*, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules," *J Am Chem Soc*, vol. 118(9), pp. 2309–2309, 1996.

[15] T. Harder *et al.*, "Beyond rotamers: a generative, probabilistic model of side chains in proteins," *BMC Bioinformatics*, vol. 11, pp. 306–319, 2010.

[16] H. Kuboniwa, N. Tjandra, S. Grzesiek, H. Ren, C. B. Klee, and A. Bax, "Solution structure of calciumfree calmodulin," *Nat Struct Biol*, vol. 2(9), pp. 768–776, 1995.