Research funded under the flagship Project “InterOmics” (PB.P05)

3D Chromatin structure estimation from Chromosome Conformation Capture data

Claudia Caudai
ISTI - CNR Pisa

BITS 2017

Research funded under the flagship Project “InterOmics” (PB.P05)

ChromStruct: method description

- Multiscale Approach
- Model evolution with quaternions
- Score function
- Recursive method with Monte Carlo algorithm

Results

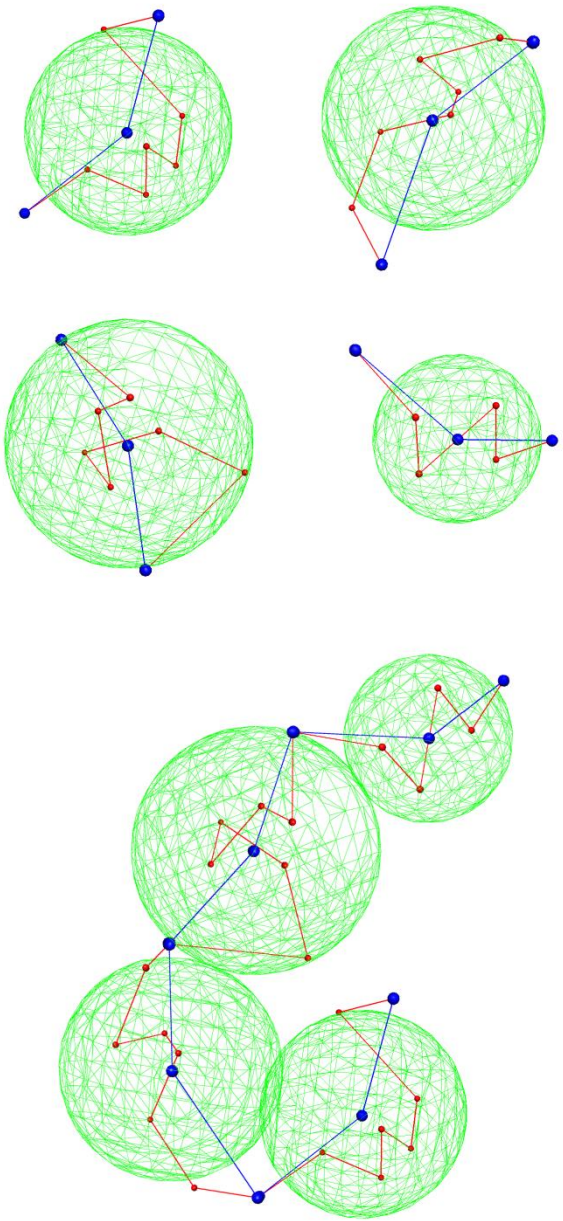
- Experiments on real Hi-C data
- Validation of results
- Comparison with TADbit

Multiscale Approach

Research funded under the flagship Project “InterOmics” (PB.P05)

Chromatin fibre is modeled as a **bead-chain**.

- The chain is divided into a number of **segments** (corrisponding to TADs) that can be treated **in parallel**.
- The procedure can be repeated **recursively** at different scales.



Experimental data are affected by **bias** derived by laboratory techniques

↓

In our data-fit function we use only contact frequencies higher than a certain treshold.

Deterministic **translation** of frequencies into distances leads to geometrical inconsistencies

↓

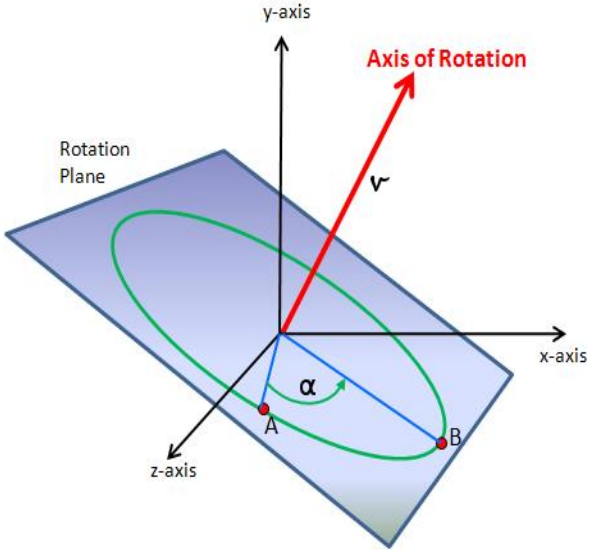
We directly introduce in our data-fit function contact frequencies, avoiding translation into distances

Model Evolution with Quaternions

Research funded under the flagship Project “InterOmics” (PB.P05)

Quaternions are an extension of the complex algebra that offers a number of advantages:

- avoiding **singularities** proper of Euler formalism (Gimbal lock)
- facilitating the **composition of rotations**
- allowing a continuous evolution maintaining topological **constraints**.



$$\mathbf{Q} = \{q_0 + q_1 i + q_2 j + q_3 k \mid q_0, q_1, q_2, q_3 \in \mathbb{R}\}$$

$$\mathbf{q} = q_0 + q_1 i + q_2 j + q_3 k = (q_0, q_1, q_2, q_3)$$

$$\bar{\mathbf{q}} = (q_0, -q_1, -q_2, -q_3)$$

$$\| \mathbf{q} \| = \sqrt{q\bar{q}} = \sqrt{q_0^2 + q_2^2 + q_2^2 + q_3^2}$$

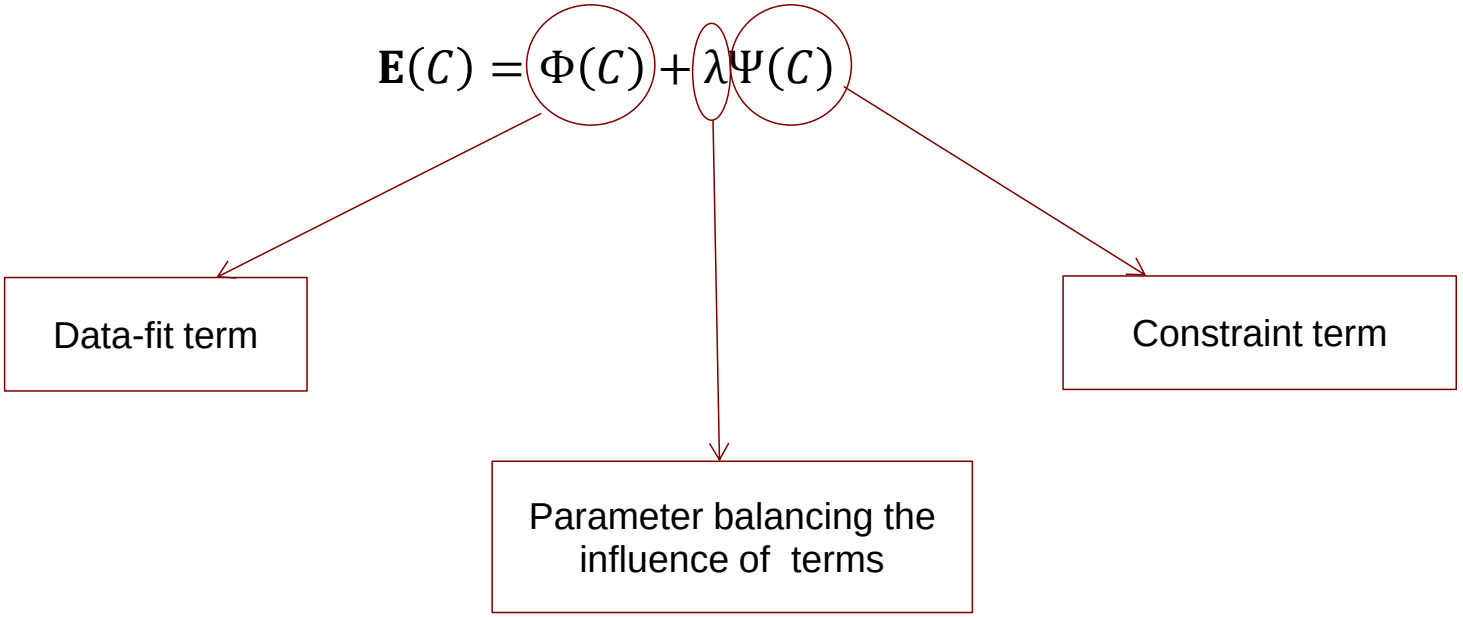
Score Function

Research funded under the flagship Project “InterOmics” (PB.P05)

To Build our energy function we assume that the bead pairs characterised by contact numbers above a certain threshold are likely to be close, whereas we do not say anything on the pairs whose contacts are below that threshold.

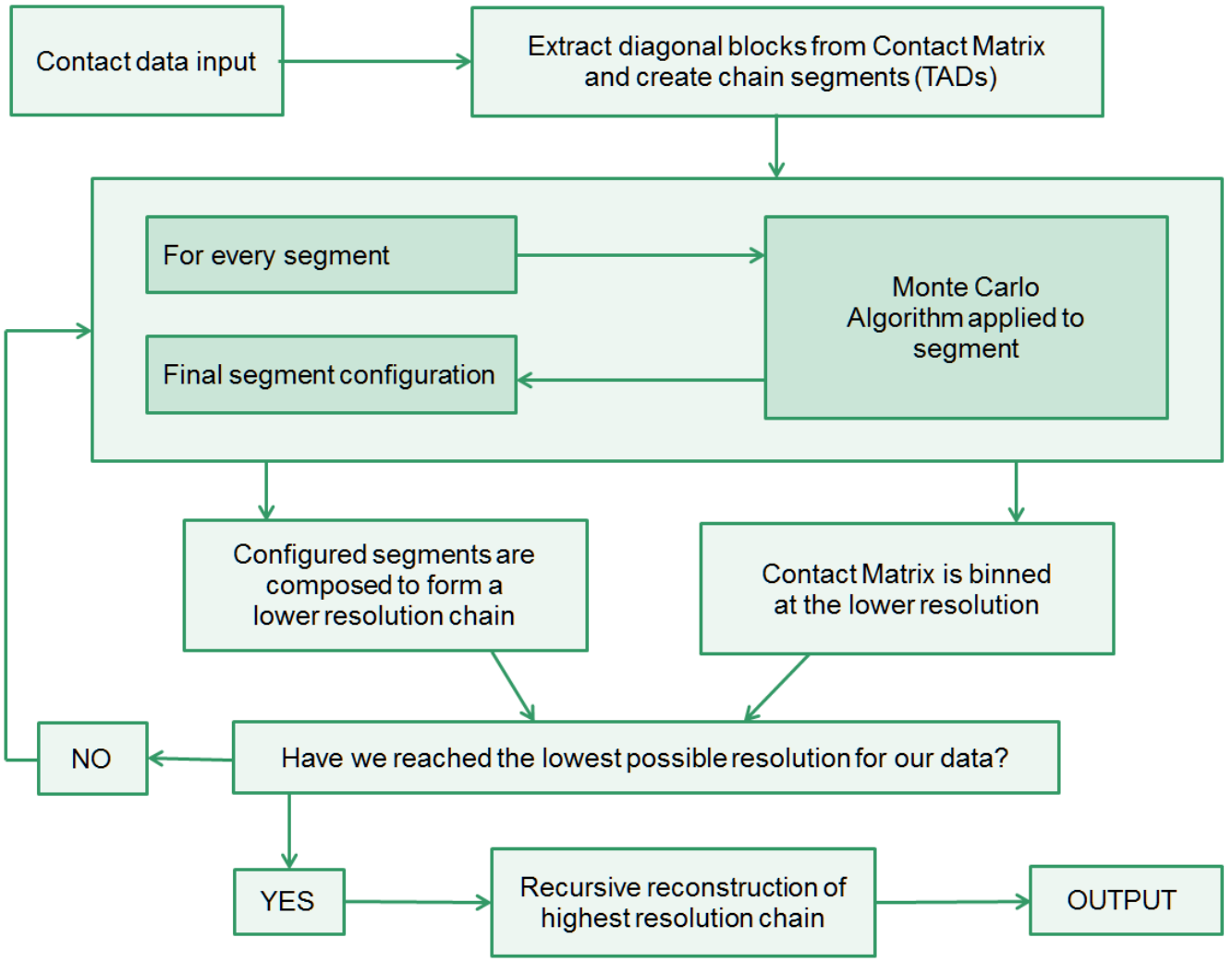
The energy of the configuration *C* is expressed by the following formula:

$$\mathbf{E}(C) = \Phi(C) + \lambda\Psi(C)$$



Recursive Method

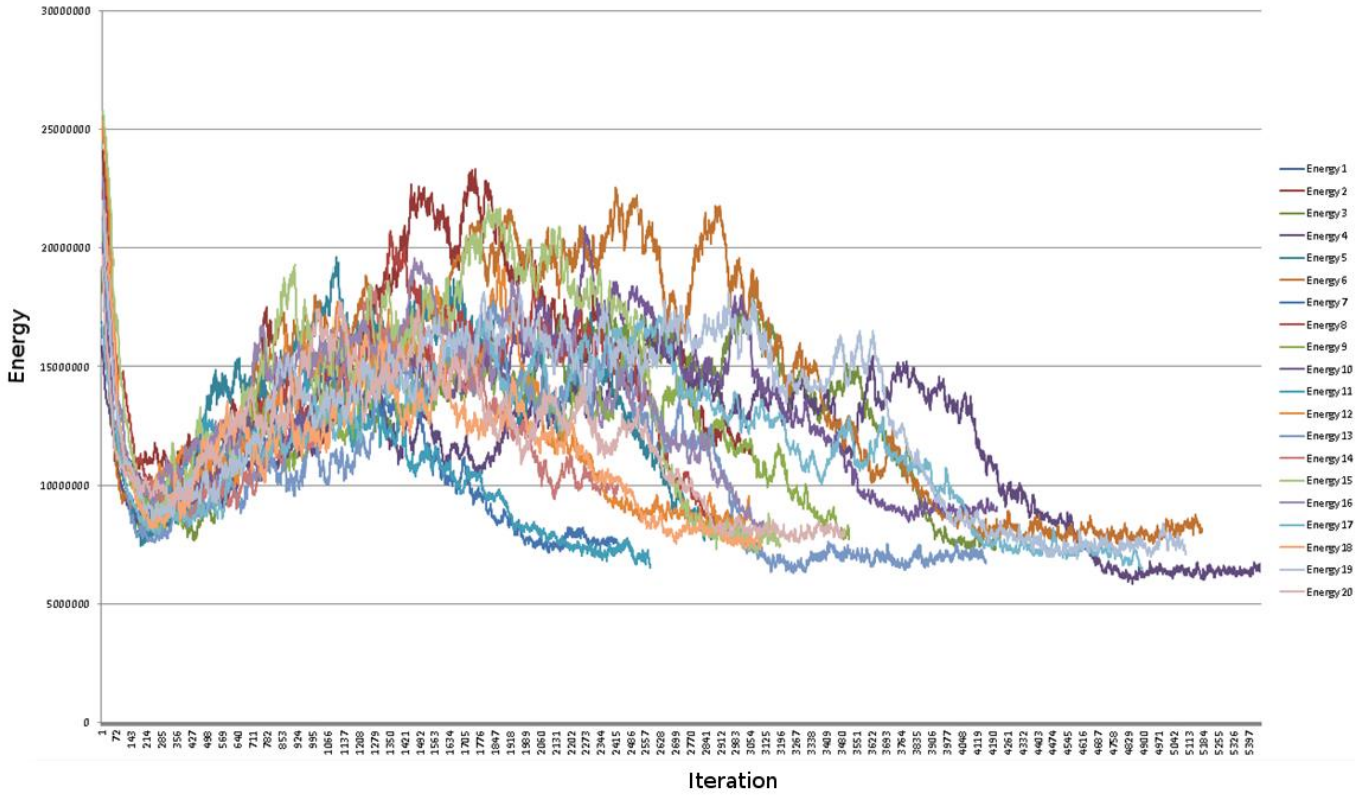
Research funded under the flagship Project “InterOmics” (PB.P05)



Simulated Annealing

Research funded under the flagship Project “InterOmics” (PB.P05)

Hi-C data are derived from millions of cells. **Simulated Annealing** allows us to widely explore the solution space.



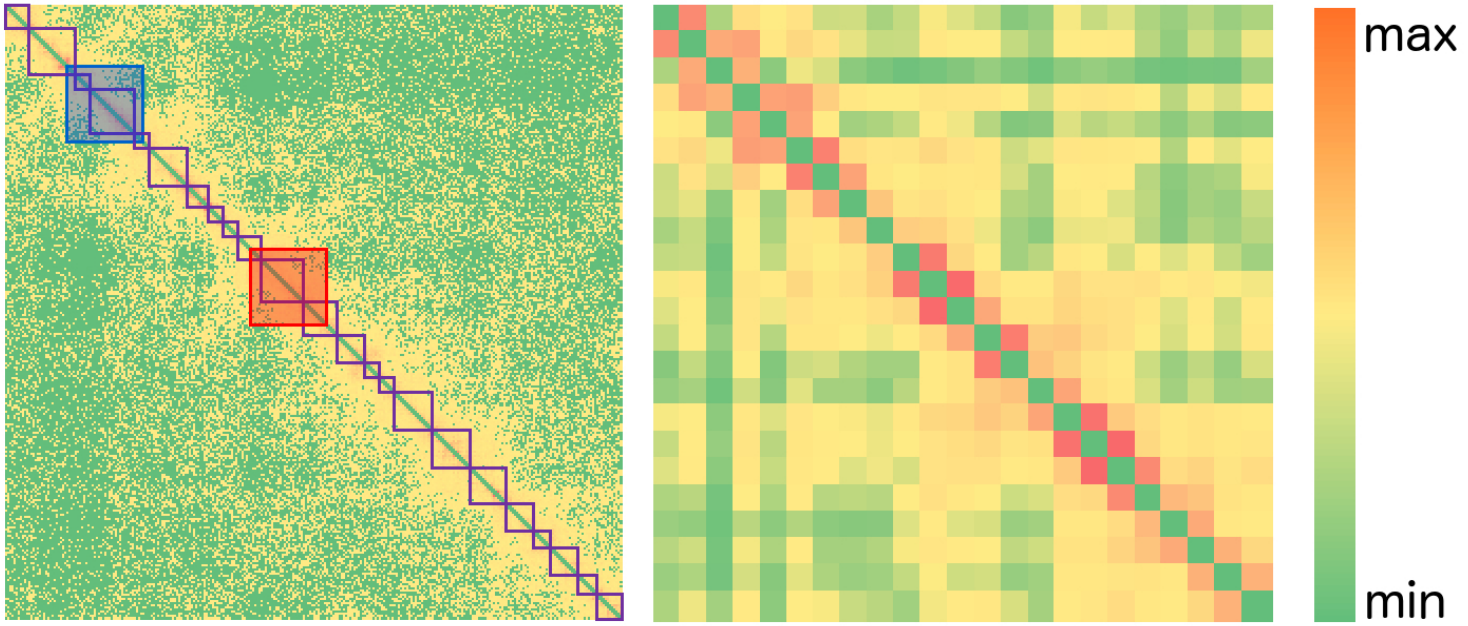
[Implementation in Python 2.7. CPU time on a 32 Core ~ 2 hours]

Tests Against Real Hi-C Data

Research funded under the flagship Project “InterOmics” (PB.P05)

ChromStruct was tested against real Hi-C data from human lymphoblastoid cells, chromosome 1, q range [150.28 Mbp, 179.44 Mbp] **Lieberman-Aiden *et al.* (2009).**

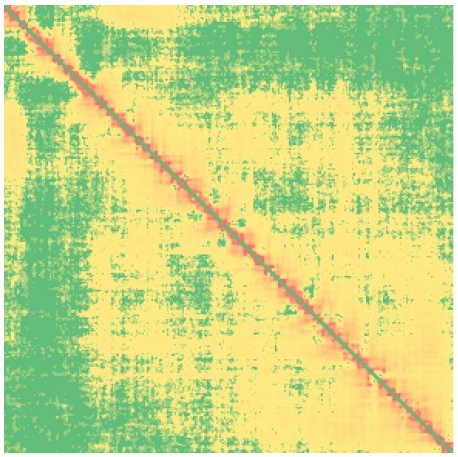
- **Higher resolution** → 100kb
- **Lower resolution** → 23 Topological Domains (7 ~ 21 Mb)

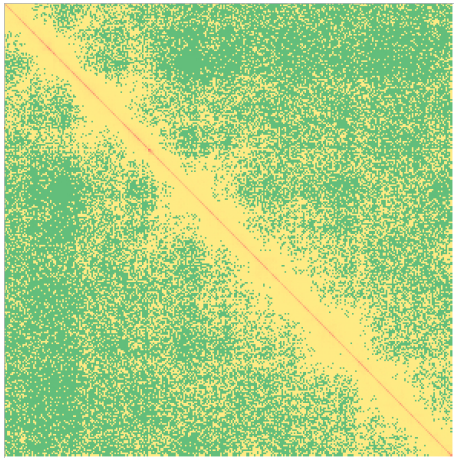


Blocks in blue and red are related, respectively, to a high-expression and a low-expression region.

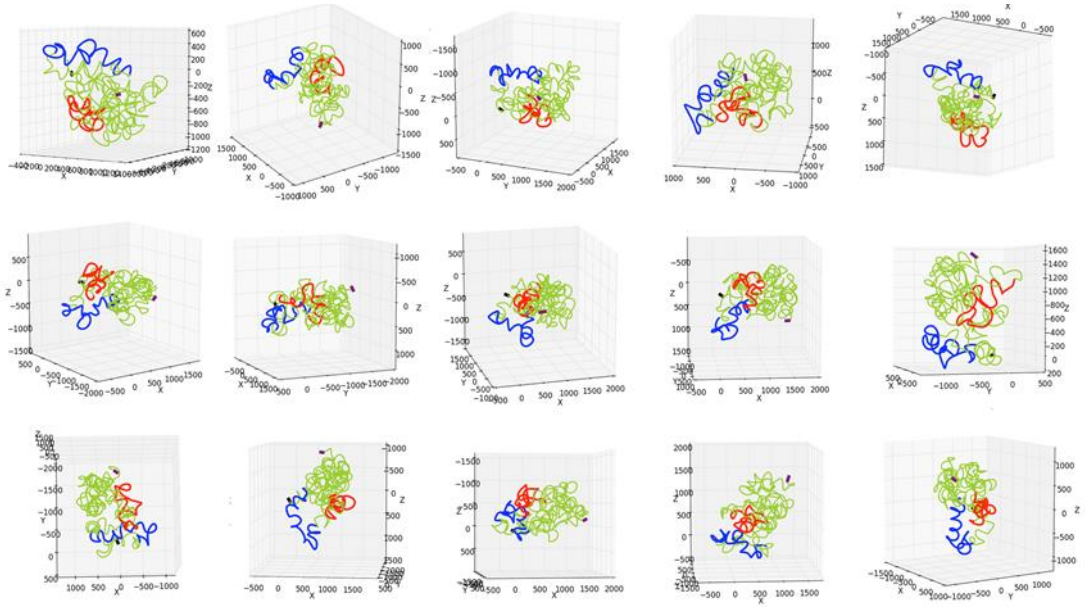
Tests Against Real Hi-C Data

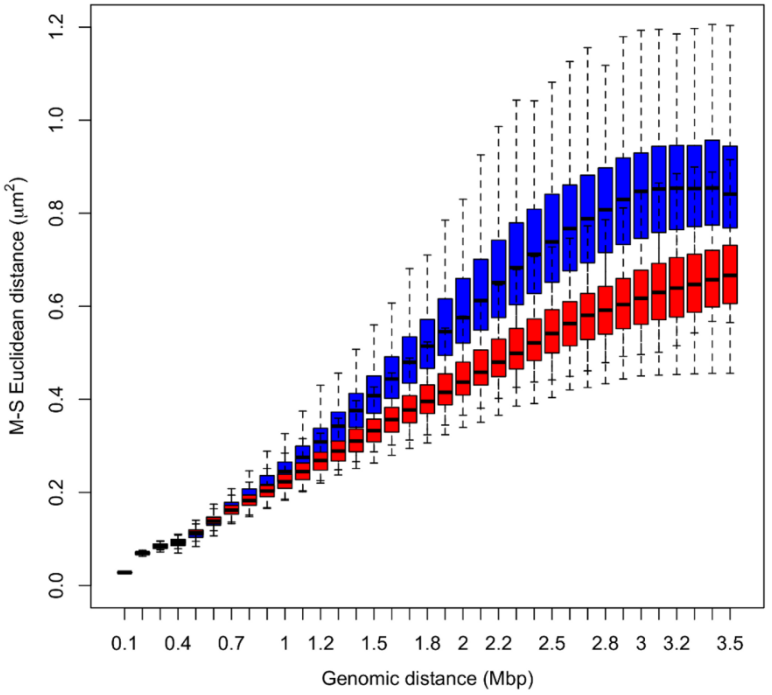
Research funded under the flagship Project “InterOmics” (PB.P05)





- Synthetic contact matrix (on the left) built from 200 configurations, compared with input contact matrix
- Boxplots of M-S Euclidean vs. Genomic distance, obtained from all our 200 solutions for the identified highly-expressed (blue) and poorly expressed (red) regions.





Comparison with TADbit

Research funded under the flagship Project “InterOmics” (PB.P05)

ChromStruct and TADbit have been compared on Hi-C data of Caulobacter Crescentus CB15 [GEO GSE45966]. Resolution 10 kb.

Problem of BIAS:

- TADbit uses **ICE** normalization method **[Imkaev *et al.* (2012)]**
- Chromstruct doesn’t use any normalization method, but a filtering technique on contact frequency matrices which makes it robust agaist biases.

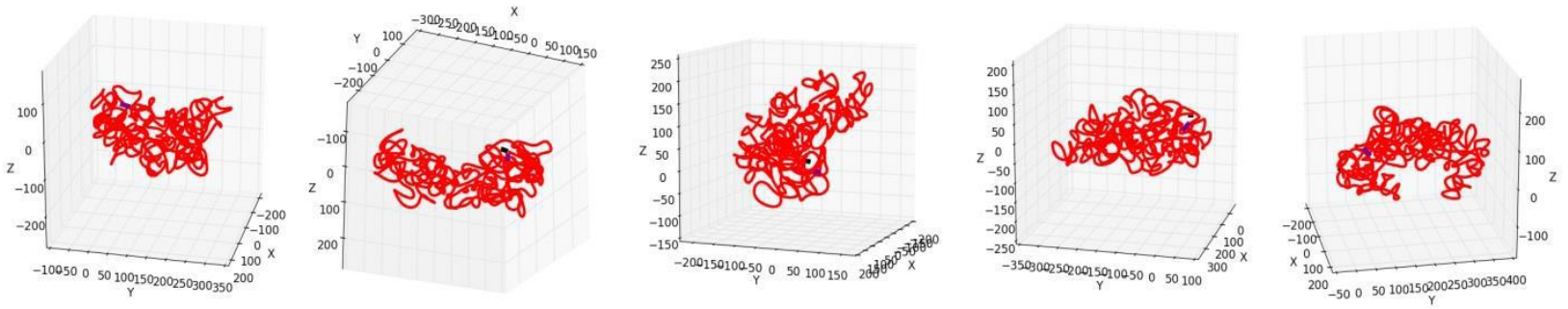
Exoperiments:

- 100 configurations with ChromStruct on raw data
- 100 configurations with ChromStruct on normalized data (ICE)
- 100 configurationds with TADbiton normalized data (ICE)

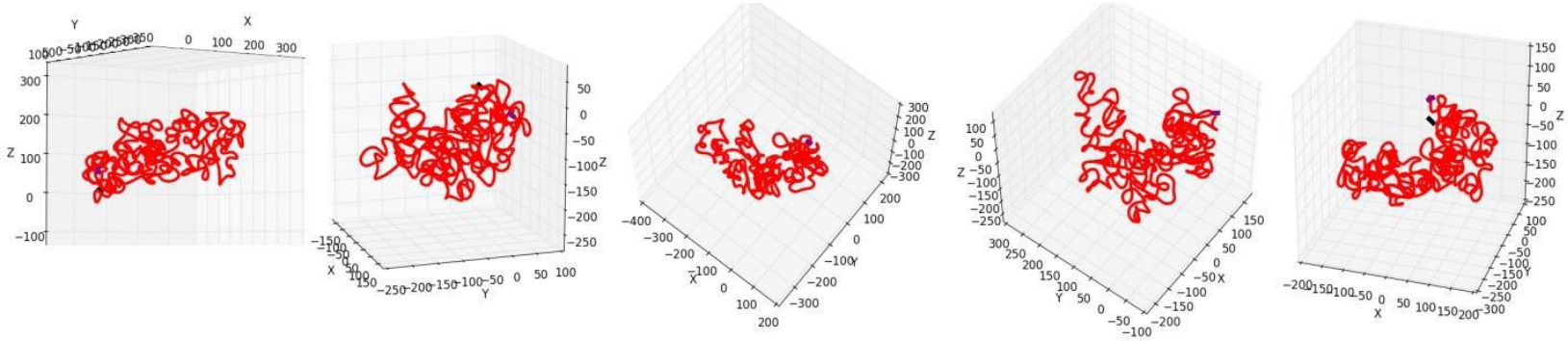
Comparison with TADbit

Research funded under the flagship Project “InterOmics” (PB.P05)

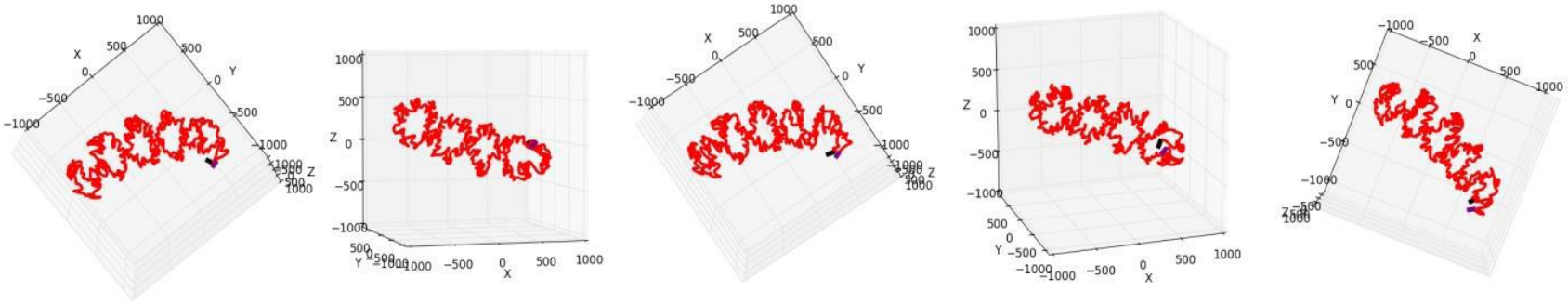
ChromStruct on raw data:



ChromStruct on normalized data:



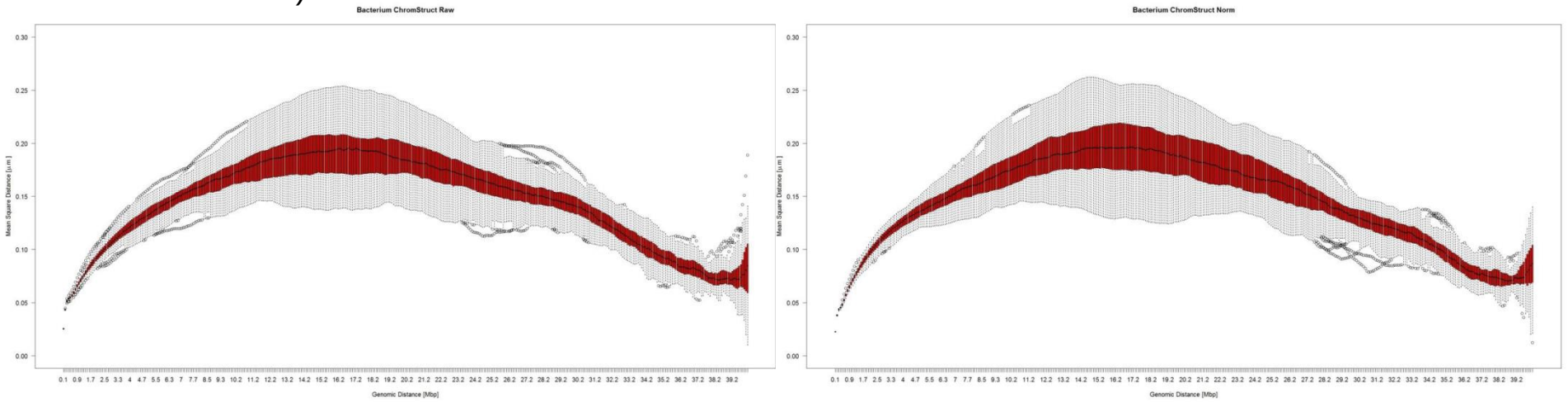
TADbit:



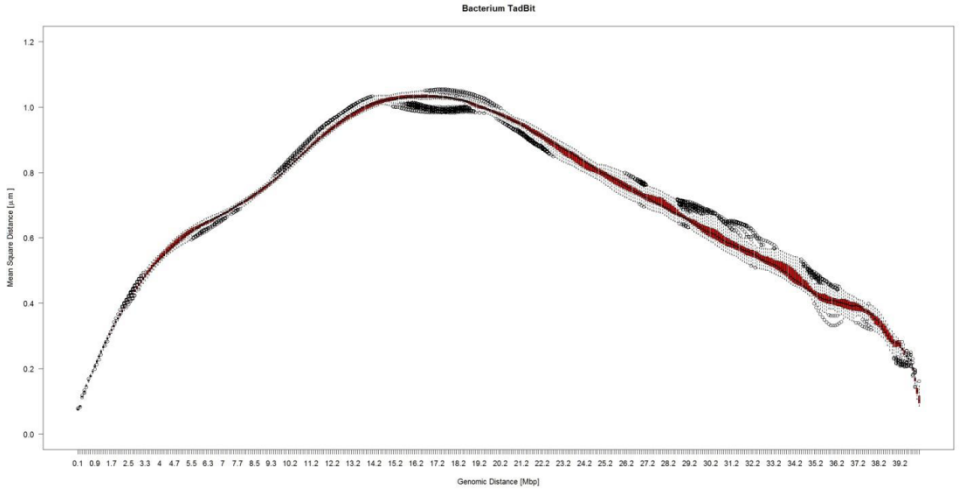
Comparison with TADbit

Research funded under the flagship Project “InterOmics” (PB.P05)

Chromstruct: boxplots for 100 configurations on raw data and 100 on normalized data (Euclidean vs. Genomic distance)



TADbit: boxplots for 100 configurations (Euclidean vs. Genomic distance)



- Chromstruct is **robust against biases**
- Configurations produced with ChromStruct are **more variated** than those produce by TADbit

Conclusions

Research funded under the flagship Project “InterOmics” (PB.P05)

Novelties and advantages of the method:

- Evolution with **quaternions** → decreasing computing time and avoiding singularities
- Score function not requiring **frequency-distance translation** → avoiding geometrical not consistent structures
- **Recursive structure** → method can be applied at different resolutions
- Robustness against **biases** → no need of data normalization

References:

- **Caudai, C. *et al.* (2015)** *Inferring 3d chromatin structure using a multiscale approach based on quaternions*, BMC Bioinformatics, 16: 234.
- **Dixon, J. R. *et al.* (2012)** *Topological domains in mammalian genomes identified by analysis of chromatin interactions,* Nature, 485, 376–380.
- **Lieberman-Aiden, E. *et al.* (2009)** *Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome,* Science, 326, 289–293.
- **Imakaev, M. *et al.* (2012)** *Iteratice correction of Hi-C data reveals allmarks of chromosome organization*, Nature Methods, 9(10):999-1003.

Authors Refernces and Aknowledgements

Research funded under the flagship Project “InterOmics” (PB.P05)

Authors:

- **Claudia Caudai**, CNR- ISTI, Pisa Italy claudia.caudai@isti.cnr.it
- **Emanuele Salerno**, CNR- ISTI, Pisa Italy
- **Monica Zoppè**, CNR- IFC, Pisa Italy
- **Anna Tonazzini**, CNR- ISTI, Pisa Italy

Aknowledgments:

We want to thank **Ivan Merelli** (CNR-ITB, Milano) for the valuable help in conducting experiments.

Authors Refernces and Aknowledgements

Research funded under the flagship Project “InterOmics” (PB.P05)

Any Questions?