Comparison between ChromStruct4 and TADbit

Flagship Project InterOmics – WP1 CNR-ISTI

Claudia Caudai [1], Emanuele Salerno [1], Monica Zoppè [2], and Anna Tonazzini [1]

National Research Council of Italy
[1]Institute of Information Science and Technologies, and [2]Institute of Clinical Physiology, Pisa, Italy
claudia.caudai@isti.cnr.it

Abstract In this report performances of ChromStruct4 and TADbit have been compared. These are two methods for the inference of chromatin three-dimensional conformations starting from Chromosome Conformation Capture data. TADbit and ChromStruct4 have been tested against the same data sets from real HI-C experiments. With comparative experiments, also robustness of ChromStruct4 against biases have been tested.

1 Introduction

TADbit [1] is a very popular method with good performances and with reasonable calculation times. It takes in input the HI-C contact data [2] from Illumina sequencing. This method removes biases through the ICE algorithm [3], produces normalized contact matrices and identifies topological association domains (TADs) [4]. TADbit models the chromatin fibre as a set of non-connected consecutive beads; all the beads have the same radius and are subject to constraints. The method identifies, in an empiric way, three different distances of balance: for consecutive beads, for couples of beads whose contact frequencies exceed a certain threshold (they are often in contact so they are supposed to be close to each other) and for couples of beads whose contact frequencies do not exceed a certain threshold (they are rarely in contact, so they are supposed to be far from each other). TADbit produces many three-dimensional conformations by optimizing harmonic functions associated with Euclidean distances between pairs of beads, using as equilibrium distances the constraints previously described. Finally, the method selects configurations best fitting the input data through a Monte Carlo method.
ChromStruct4 [5, 6] takes in input HI-C contact matrices. The method doesn't clean up data because it uses a data filtering method which makes it robust against biases. ChromStruct4 identifies topological domains (TADs) using its own algorithm and models the chromatin fibre as a bead-chain in which consecutive beads are connected to each other. The system is treated as a kinematic

chain and motion is performed by quaternions. HI-C data derive from millions of cells, so ChromStruct4 is aimed to sample the space of possible solutions. At every run, this method produces a three-dimensional chromatin conformation through a Simulated Annealing method, adopting a scoring function oriented to rewarding fit with input data and discouraging geometrical constraint violations.

2 Tests against Hi-C real data

Both methods have been tested against the following data sets:

- Long arm of human Chromosome 1: from $150.3Mb$ to $179.4Mb$ (binned at $100kb$) [GEO: GSE18199] [2]
- Caulobacter Crescentus CB15 (binned at $10kb$) [GEO: GSE45966] [3]

TADbit takes in input fastq sequencing data and produces contact matrices, both raw and normalized with ICE. The raw matrix produced by TADbit for the long arm of Chromosmome 1 is different from the contact matrix of Lieberman-Aiden [2], even though they both derive from the same HI-C data, also the raw Caulobacter matrix produced by TADbit is different from the contact matrix of Imakaev [3], even though same HI-C data have been used.

Observing the heatmaps of contact matrices produced by TADbit (Figures 2-3 and Figures 5-6) and values inside contact matrices (Figures 7-10), it is noteworthy that contact frequencies are different before and after normalization with ICE (two or three orders of magnitude), but the differences in heatmaps are small. Probably, the heatmap MICROSOFT EXCEL's algorithm attributes to raw and normalized values the same color.
We also observe that:

- Matrices produced by TADbit for Chromosome 1 are more concentrated around the main diagonal than those of Lieberman-Aiden
- Matrices produced by TADbit for the Caulobacter Crescentus have many rows and columns full of zeroes.

Probably, the differences between data produced by TADbit and Lieberman-Aiden and between data produced by TADbit and Imakaev depend on differences in the methods used in building the contact matrices from sequencing data.

3 Comparison of conformations produced by TADbit and ChromStruct4

In order to compare the two methods, the contact matrices produced by TADbit, both normalized and raw, have been used. Aims of this comparison are:

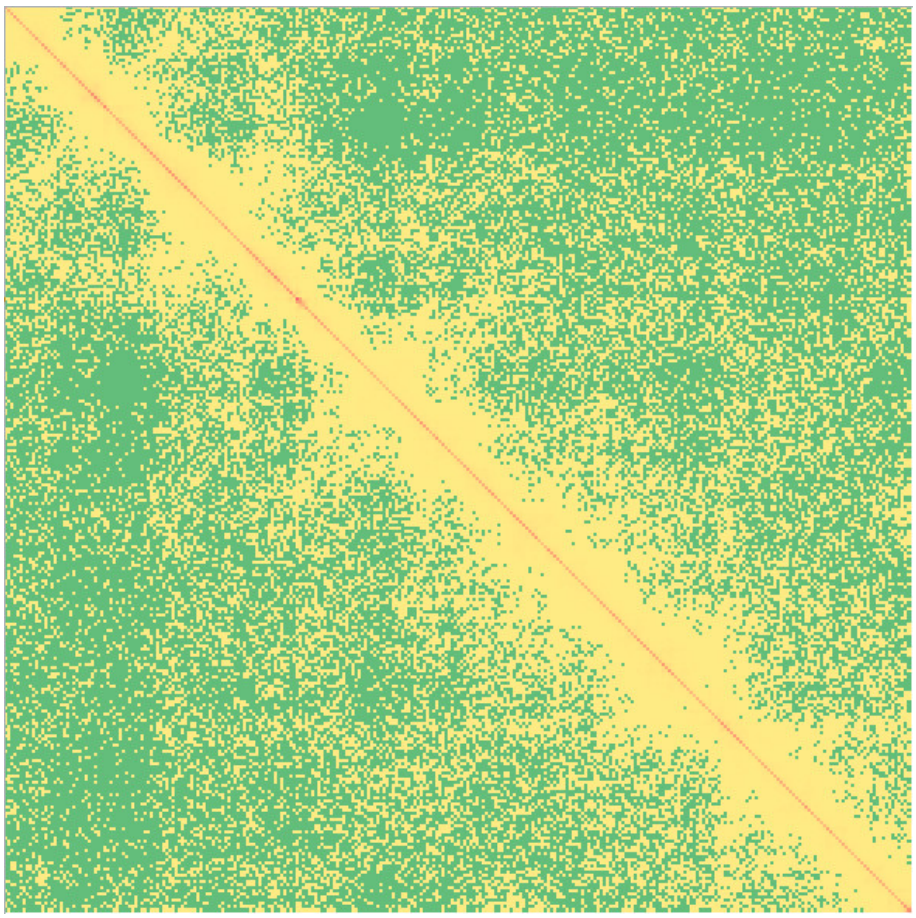


Figure1. Heatmap of the contact matrix of the long arm of Chromosome 1 produced by Lieberman-Aiden on HI-C data.

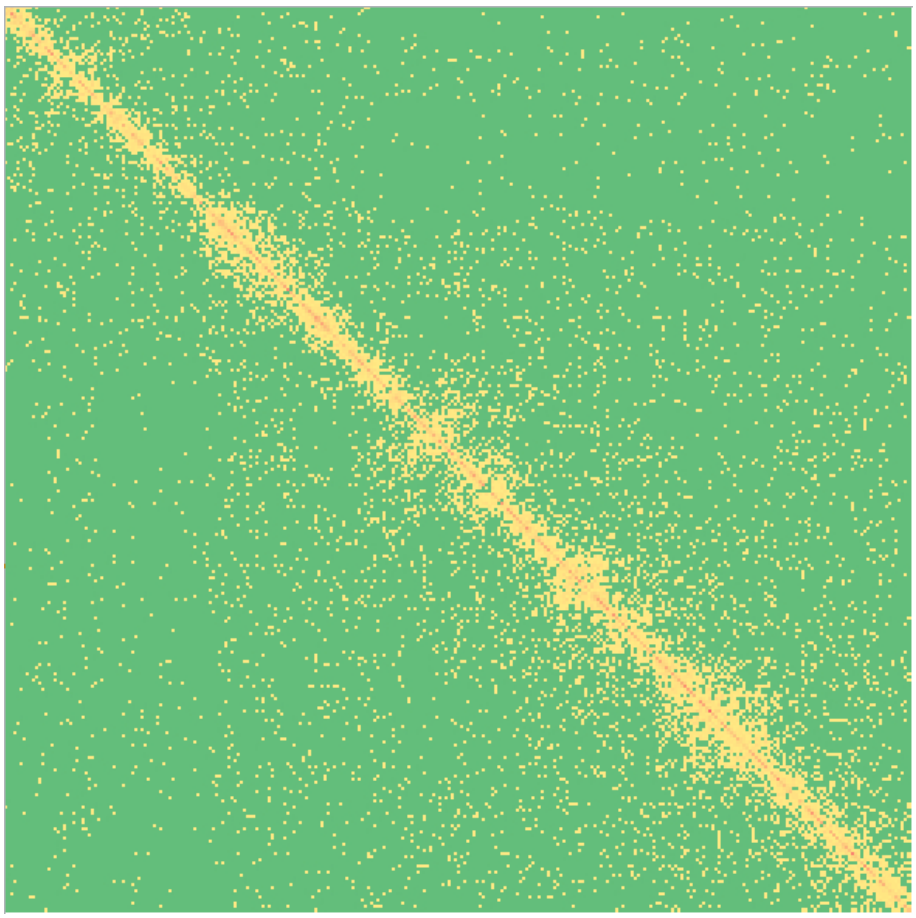


Figure2. Heatmap of the raw contact matrix of the long arm of Chromosome 1 produced by TADbit.

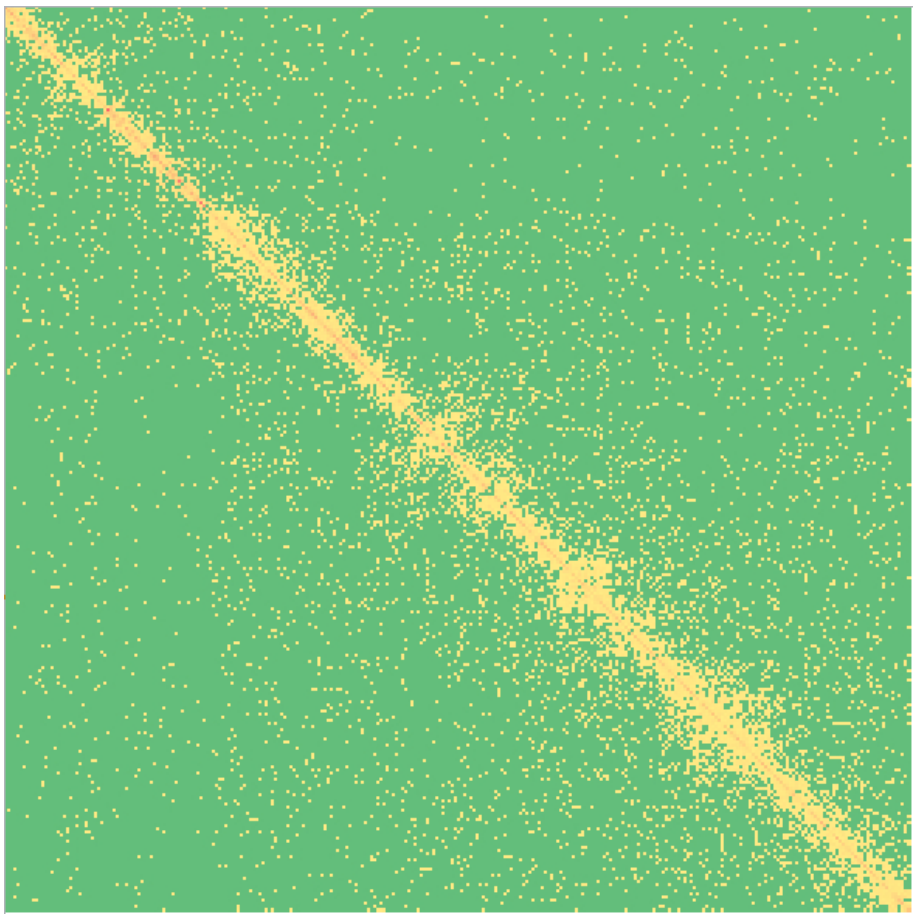


Figure3. Heatmap of the normalized (ICE) contact matrix of the long arm of Chromosome 1 produced by TADbit.

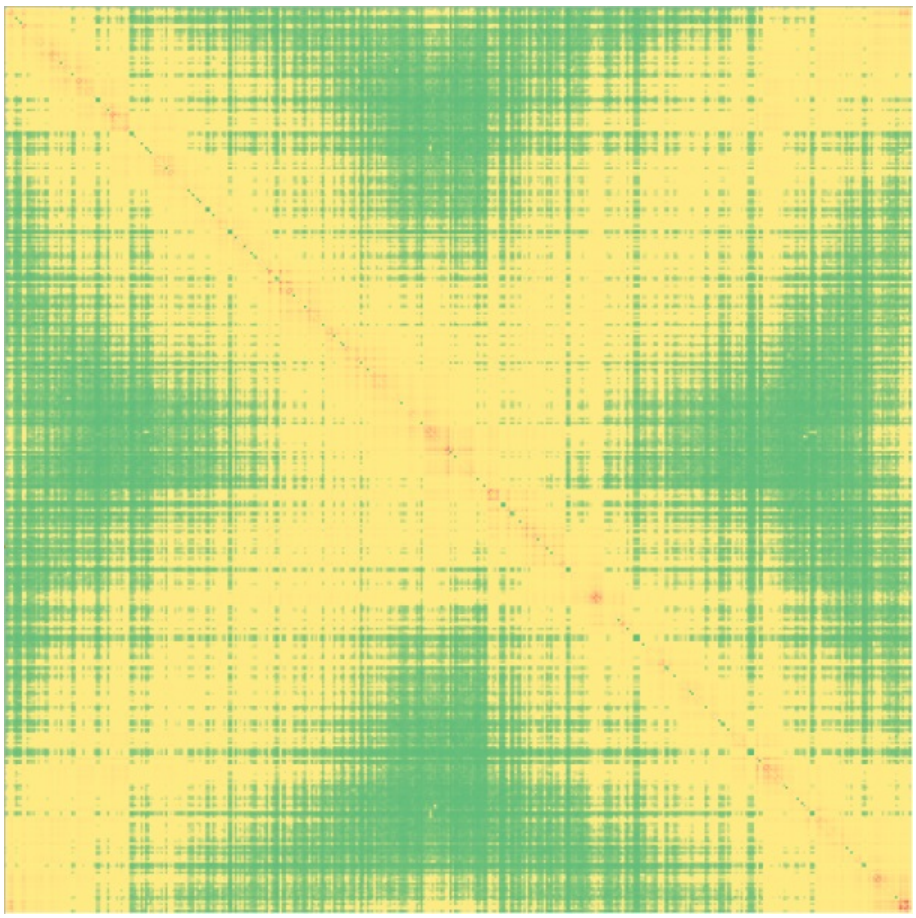


Figure4. Heatmap of the contact matrix of Caulobacter Crescentus produced by Imakaev on HI-C data.

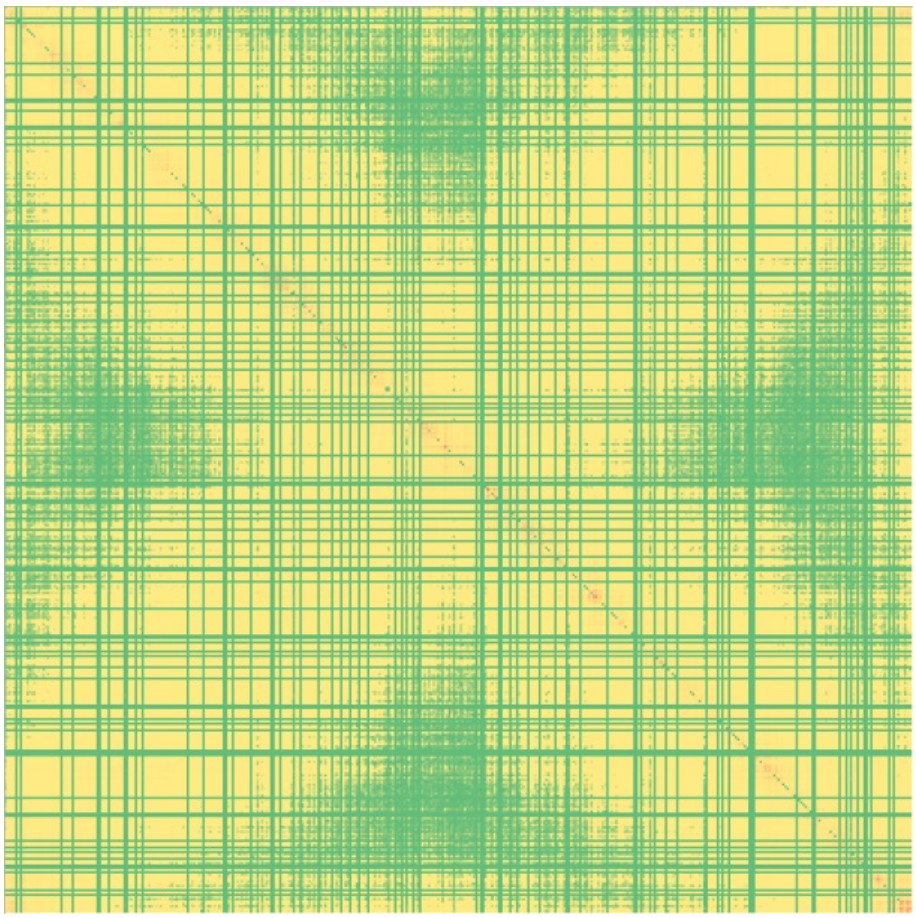


Figure5. Heatmap of the raw contact matrix of Caulobacter Crescentus produced by TADbit.

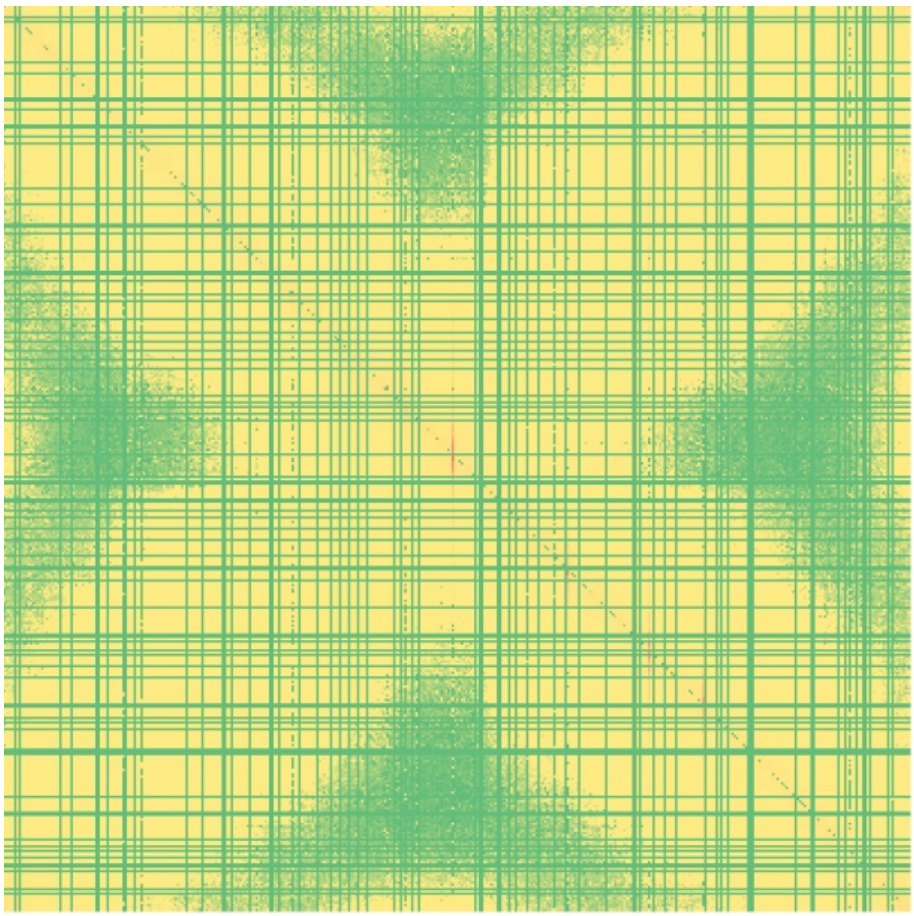
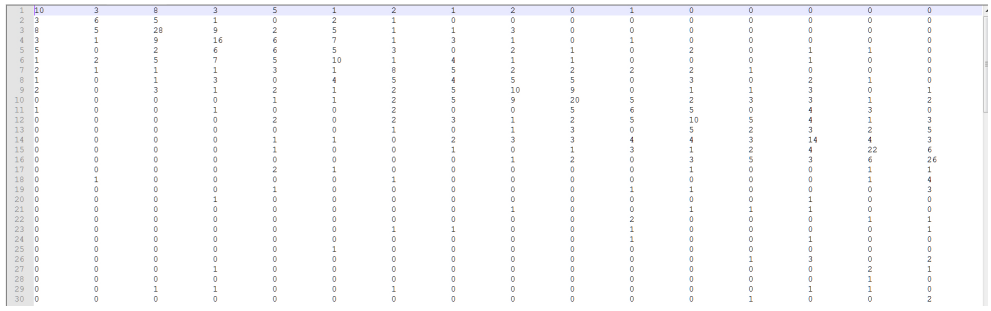
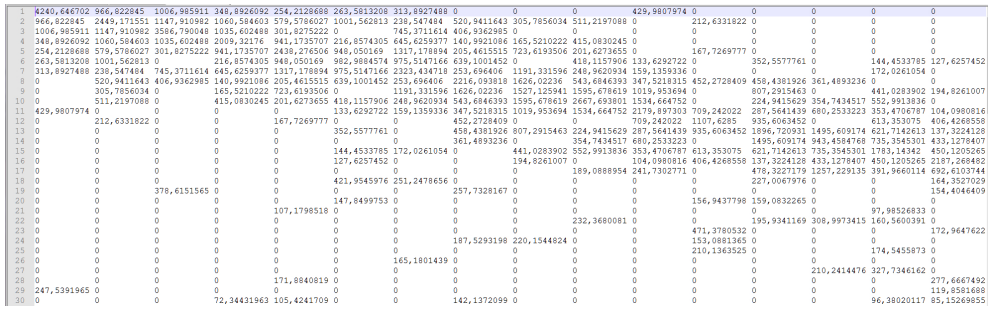


Figure6. Heatmap of the normalized (ICE) contact matrix of Caulobacter Crescentus produced by TADbit.



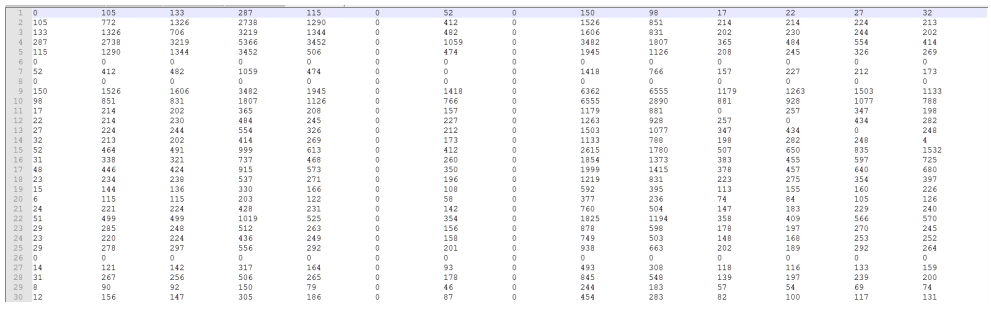
1	10	3	8	3	5	1	2	1	2	0	1	0	0	0	0	0
2	3	6	5	1	0	2	1	0	0	0	0	0	0	0	0	0
3	8	5	28	9	2	5	1	1	3	0	0	0	0	0	0	0
4	3	1	9	16	6	7	1	3	1	0	1	0	0	0	0	0
5	5	0	2	6	6	5	3	0	2	1	0	2	0	1	1	0
6	1	2	5	7	5	10	1	4	1	1	0	0	0	1	0	0
7	2	1	1	1	3	1	8	5	2	2	2	2	1	0	0	0
8	1	0	1	3	0	4	5	4	5	5	0	3	0	2	1	0
9	2	0	3	1	2	1	2	5	10	9	0	1	1	3	0	1
10	0	0	0	0	1	1	2	5	9	20	5	2	3	3	1	2
11	1	0	0	1	0	0	2	0	0	5	6	5	0	4	3	0
12	0	0	0	0	2	0	2	3	1	2	5	10	5	4	1	3
13	0	0	0	0	0	0	1	0	1	3	0	5	2	3	2	5
14	0	0	0	0	1	1	0	2	3	3	4	4	3	14	4	3
15	0	0	0	0	1	0	0	1	0	1	3	1	2	4	22	6
16	0	0	0	0	0	0	0	0	1	2	0	3	5	3	6	26
17	0	0	0	0	2	1	0	0	0	0	0	1	0	0	1	1
18	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	4
19	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	3
20	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
21	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0
22	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	1
23	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	1
24	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
25	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	1	3	0	2
27	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	1
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
29	0	0	1	1	0	0	1	0	0	0	0	0	0	1	1	0
30	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2

Figure7. Contact matrix of raw data of the long arm of Chromosome 1 produced by TADbit.



1	4240,646702	966,822845	1006,985911	348,8926092	254,2128688	263,5813208	313,8927488	0	0	0	429,9807974	0	0	0	0	0
2	966,822845	2449,171551	1147,910982	1060,584603	579,5786027	1001,562813	238,547484	520,9411643	305,7856034	511,2197088	0	212,6331822	0	0	0	0
3	1006,985911	1147,910982	3586,790048	1035,602488	301,8275222	0	745,3711614	406,9362985	0	0	0	0	0	0	0	0
4	348,8926092	1060,584603	1035,602488	2009,32176	941,1735707	216,8574305	645,6259377	140,9921086	165,5210222	415,0830245	0	0	0	0	0	0
5	254,2128688	579,5786027	301,8275222	941,1735707	2438,276506	946,050169	1317,178894	205,4615515	723,6193506	201,6273655	0	167,7269777	0	0	0	0
6	263,5813208	1001,562813	0	216,8574305	948,050169	982,9884574	975,5147166	639,1001452	0	418,1157906	133,6292722	0	352,5577761	0	144,4533785	127,6257452
7	313,8927488	238,547484	745,3711614	645,6259377	1317,178894	975,5147166	2323,434718	253,696406	1191,331596	248,9620934	159,1359336	0	0	0	172,0261054	0
8	0	520,9411643	406,9362985	140,9921086	205,4615515	639,1001452	253,696406	2216,093818	1626,02236	543,6846393	347,5218315	452,2728409	458,4381926	361,4893236	0	0
9	0	305,7856034	0	165,5210222	723,6193506	0	1191,331596	1626,02236	1527,125941	1595,678619	1019,953694	0	807,2915463	0	441,0283902	194,8261007
10	0	511,2197088	0	415,0830245	201,6273655	418,1157906	248,9620934	543,6846393	1595,678619	2667,693801	1534,664752	0	224,9415629	354,7434517	552,9913836	0
11	429,9807974	0	0	0	0	133,6292722	159,1359336	347,5218315	1019,953694	1534,664752	2179,897303	709,242022	287,5641439	680,2533223	353,4706787	104,0980816
12	0	212,6331822	0	0	167,7269777	0	0	452,2728409	0	0	709,242022	1107,6285	935,6063452	0	613,353075	406,4268558
13	0	0	0	0	0	352,5577761	0	458,4381926	807,2915463	224,9415629	287,5641439	935,6063452	1896,720931	1495,609174	621,7142613	137,3224128
14	0	0	0	0	0	0	0	361,4893236	0	354,7434517	680,2533223	0	1495,609174	943,4584768	735,3545301	433,1278407
15	0	0	0	0	0	144,4533785	172,0261054	0	441,0283902	552,9913836	353,4706787	613,353075	621,7142613	735,3545301	1783,14342	450,1205265
16	0	0	0	0	0	127,6257452	0	0	194,8261007	0	104,0980816	406,4268558	137,3224128	433,1278407	450,1205265	2187,268482
17	0	0	0	0	0	0	0	0	0	189,0888954	241,7302771	0	478,3227179	1257,229135	391,9660114	692,6103744
18	0	0	0	0	0	421,9545976	251,2478656	0	0	0	0	0	227,0067976	0	0	164,3527029
19	0	0	378,6151565	0	0	0	0	257,7328167	0	0	0	0	0	0	0	154,4046409
20	0	0	0	0	0	147,8499753	0	0	0	0	0	156,9437798	159,0832265	0	0	0
21	0	0	0	0	107,1798518	0	0	0	0	0	0	0	0	0	97,98526833	0
22	0	0	0	0	0	0	0	0	0	232,3680081	0	0	195,9341169	308,9973415	160,5600391	0
23	0	0	0	0	0	0	0	0	0	0	0	471,3780532	0	0	0	172,9647622
24	0	0	0	0	0	0	0	187,5293198	220,1544824	0	0	153,0881365	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	210,1363525	0	0	174,5455873	0
26	0	0	0	0	0	0	165,1801439	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	210,2414476	327,7346162	0
28	0	0	0	0	171,8840819	0	0	0	0	0	0	0	0	0	0	277,6667492
29	247,5391965	0	0	0	0	0	0	0	0	0	0	0	0	0	0	119,8581688
30	0	0	0	72,34431963	105,4241709	0	0	142,1372099	0	0	0	0	0	0	96,38020117	85,15289855

Figure8. Contact matrix of normalized (ICE) data of the long arm of Chromosome 1 produced by TADbit.



1	0	105	133	287	115	0	52	0	150	98	17	22	27	32	
2	105	772	1326	2738	1290	0	412	0	1526	851	214	214	224	213	
3	133	1326	706	3219	1344	0	482	0	1606	831	202	230	244	202	
4	287	2738	3219	5366	3452	0	1059	0	3482	1807	365	484	554	414	
5	115	1290	1344	3452	506	0	474	0	1945	1126	208	245	326	269	
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	52	412	482	1059	474	0	0	0	1418	766	157	227	212	173	
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9	150	1526	1606	3482	1945	0	1418	0	6362	6555	1179	1263	1503	1133	
10	98	851	831	1807	1126	0	766	0	6555	2890	881	928	1077	788	
11	17	214	202	365	208	0	157	0	1179	881	0	257	347	198	
12	22	214	230	484	245	0	227	0	1263	928	257	0	434	282	
13	27	224	244	554	326	0	212	0	1503	1077	347	434	0	248	
14	32	213	202	414	269	0	173	0	1133	788	198	282	248	4	
15	52	464	491	999	613	0	412	0	2615	1780	507	650	835	1532	
16	31	338	321	737	468	0	260	0	1854	1373	383	455	597	725	
17	48	446	424	915	573	0	350	0	1999	1415	378	457	640	680	
18	23	234	238	537	271	0	196	0	1219	831	223	275	354	397	
19	15	144	136	330	166	0	108	0	592	395	113	155	160	226	
20	6	115	115	203	122	0	58	0	377	236	74	84	105	126	
21	24	221	224	428	231	0	142	0	760	504	147	183	229	240	
22	51	499	499	1019	525	0	354	0	1825	1194	358	409	566	570	
23	29	285	248	512	263	0	156	0	878	598	178	197	270	245	
24	23	220	224	436	249	0	158	0	749	503	148	168	253	252	
25	29	278	297	556	292	0	201	0	938	663	202	189	292	264	
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
27	14	121	142	317	164	0	93	0	493	308	118	116	133	159	
28	31	267	256	506	265	0	178	0	845	548	139	197	239	200	
29	8	90	92	150	79	0	46	0	244	183	57	54	69	74	
30	12	156	147	305	186	0	87	0	454	283	82	100	117	131	

Figure9. Contact matrix of raw data of Caulobacter Crescentus produced by TADbit.

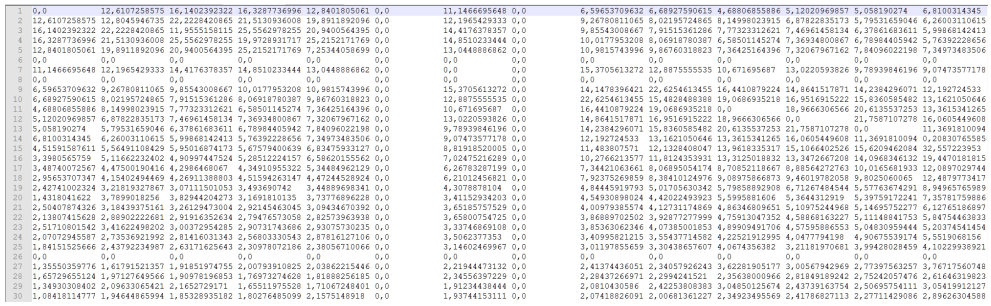


Figure10. Contact matrix of normalized (ICE) data of Caulobacter Crescentus produced by TADbit.

1. Verifying that the outputs of ChromStruct4 do not depend on data normalization, because ChromStruct4 uses a filtering method that makes it robust against biases.
2. Comparing final configurations produced by TADbit and ChromStruct4 from the same inputs. Agreement of results would be a good validation, because the two methods are based on different approaches. If the final outputs are similar, the features characterizing such configurations could actually mark biological characteristics captured by both methods.

The following results have been produced:

- ChromStruct4: 100 configurations of Chromosome 1 starting from the raw matrix
- ChromStruct4: 100 configurations of Chromosome 1 starting from the normalized matrix
- TADbit: 100 configurations of Chromosome 1 starting from the normalized matrix
- ChromStruct4: 100 configurations of Caulobacter starting from the raw matrix
- ChromStruct4: 100 configurations of Caulobacter starting from the normalized matrix
- TADbit: 100 configurations of Caulobacter starting from the normalized matrix

Figures 11-16 demonstrate how the configurations produced by ChromStruct4 are more varied in shape than those produced by TADbit, which are very homogeneous. Conformations of both Chromosome 1 and Caulobacter produced by TADbit are very similar to each other, this behavior may be due to the algorithm leaving few degrees of freedom or to the biases given by harmonics governing TADbit's evolution.

Figures 17-22 show the boxplots of Euclidean-distance vs genomic-distance for all output sets. These graphs help us to understand the variability of the

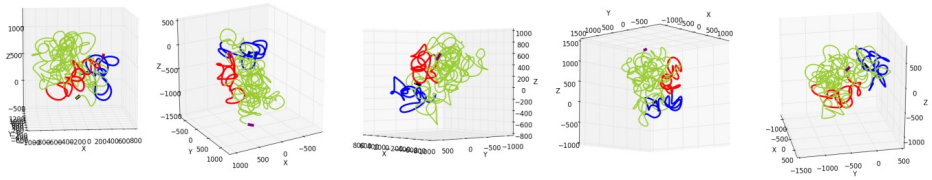


Figure11. Configurations of Chromosome 1 produced by ChromStruct4 from the raw contact matrix.

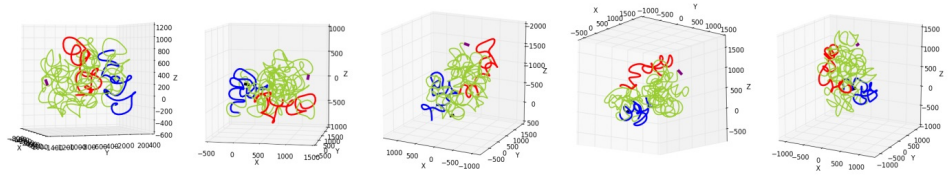


Figure12. Configurations of Chromosome 1 produced by ChromStruct4 from the normalized contact matrix.

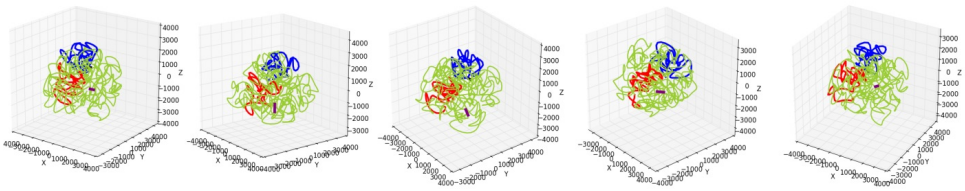


Figure13. Configurations of Chromosome 1 produced by TADbit from the normalized contact matrix.

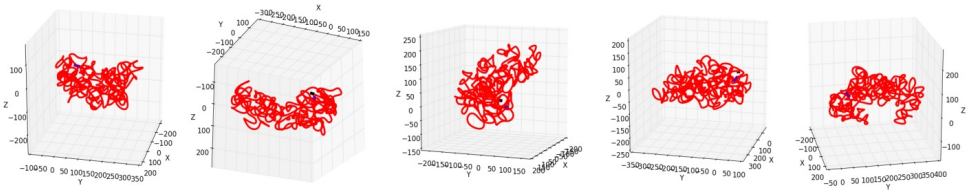


Figure14. Configurations of Caulobacter produced by ChromStruct4 from the raw contact matrix.

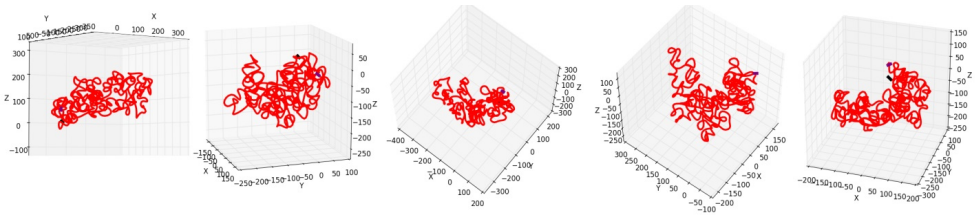


Figure15. Configurations of Caulobacter produced by ChromStruct4 from the normalized contact matrix.

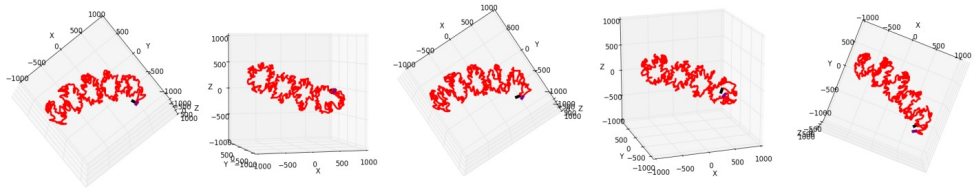


Figure16. Configurations of Caulobacter produced by ChromStruct4 from the normalized contact matrix.

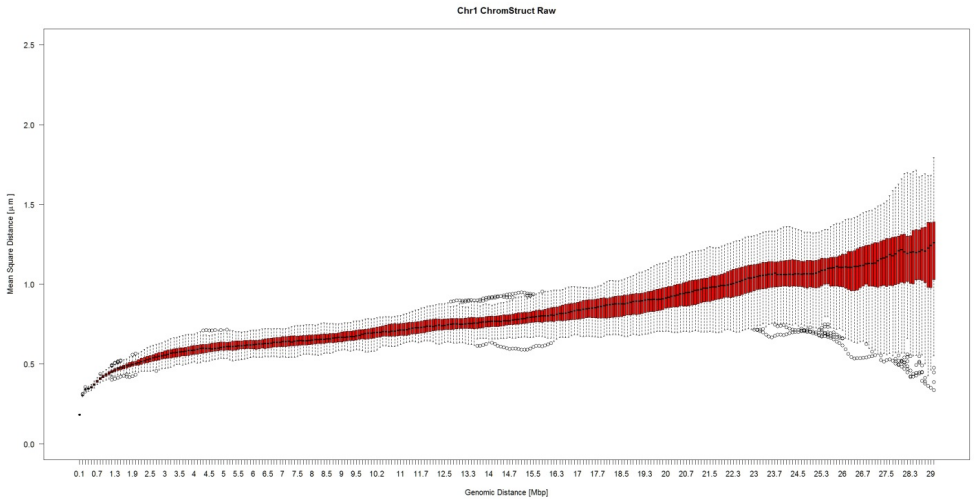


Figure17. Boxplot of 100 conformations produced by ChromStruct4 on raw data of Chromosome 1.

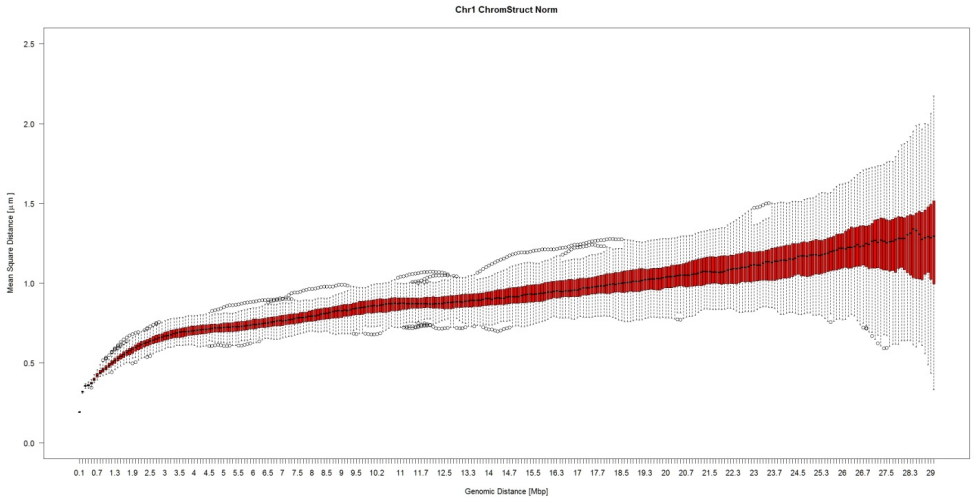


Figure18. Boxplot of 100 conformations produced by ChromStruct4 on normalized data of Chromosome 1.

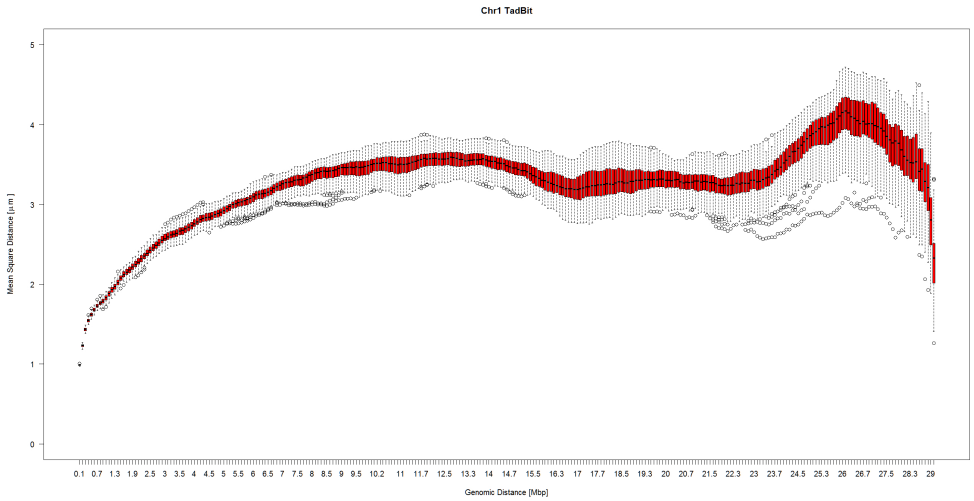


Figure19. Boxplot of 100 conformations produced by TADbit on normalized data of Chromosome 1.

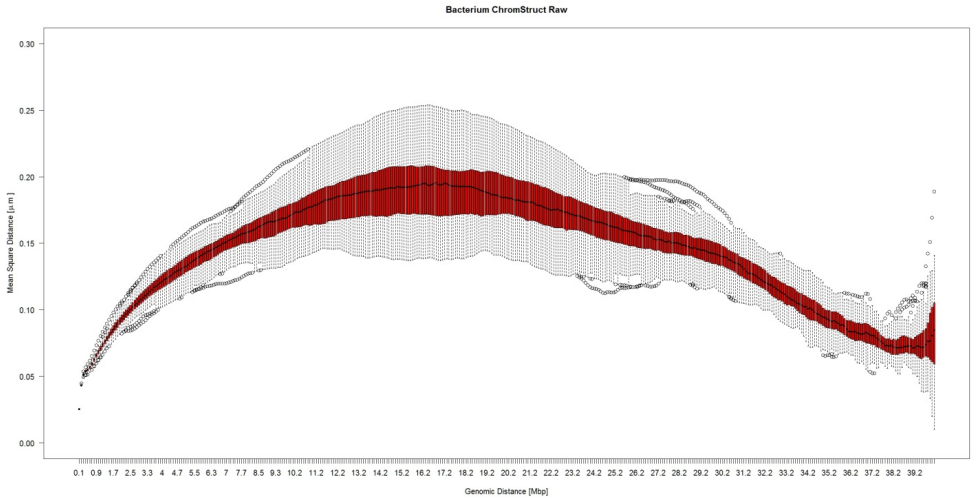


Figure20. Boxplot of 100 conformations produced by ChromStruct4 on raw data of Caulobacter.

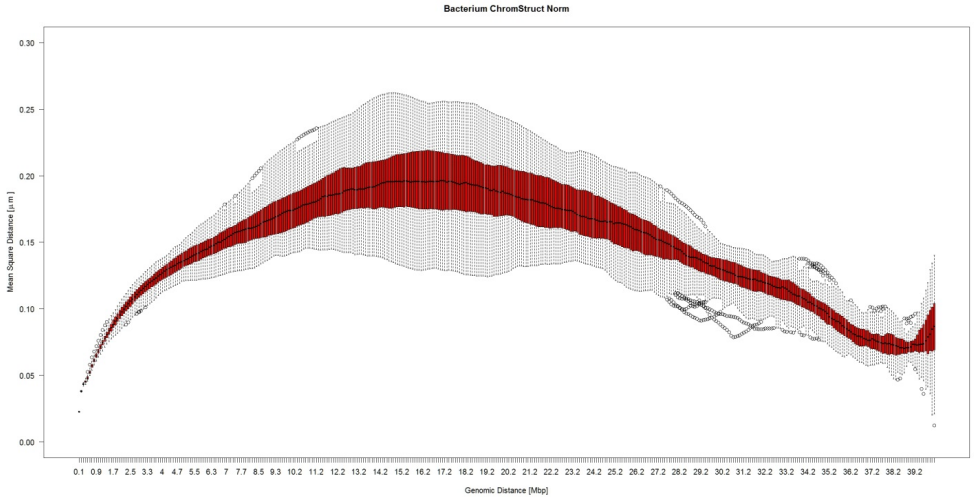


Figure21. Boxplot of 100 conformations produced by ChromStruct4 on normalized data of Caulobacter.

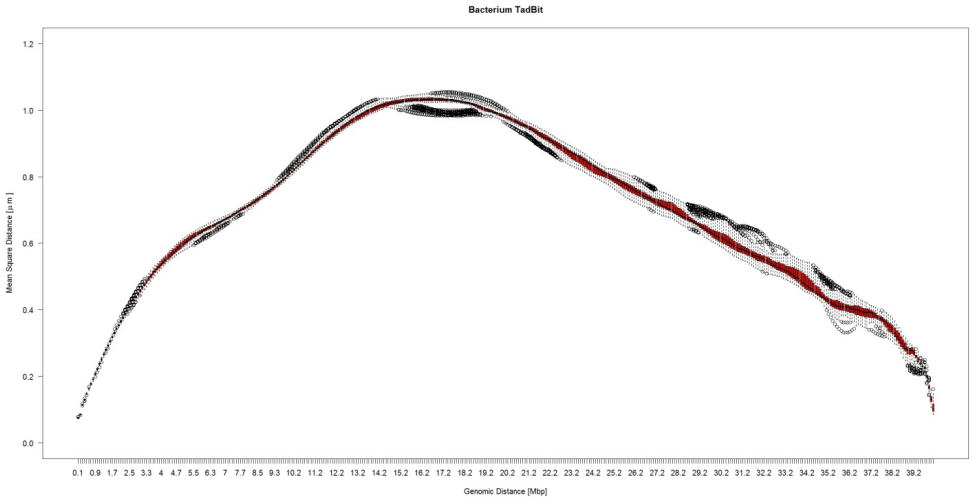


Figure22. Boxplot of 100 conformations produced by TADbit on normalized data of Caulobacter.

solutions with respect to their genomic organization. As can be seen, the configurations of Caulobacter and Chromosome 1 produced by TADbit are much less variable than those produced by ChromStruct4 (boxplots are less expanded). Other important observations can be made about size:

- The maximum size of the portion of Chromosome 1 [150.3Mb,179.4Mb] modelled by ChromStruct4 is about 2.5 microns, while the conformations produced by TADbit on the same data reach nearly 5 microns.
- The size of the Caulobacter conformations produced by ChromStruct4 are smaller, never exceeding 0.3 microns, than those produced by TADbit, always more than 1 micron.

Synthetic contact matrices have been produced for both ChromStruct4 and TADbit. Threshold of contact between two beads has been assumed as the sum of their radii (calculated by ChromStruct4) multiplied by 1.2.

Figures 23-28 show that the conformations produced by TADbit are much less varied than those produced by ChromStruct4, the contacts are infrequently scattered and almost always concern the same bin pairs. Moreover, for Caulobacter, the red areas (i.e. high frequencies) are concentrated around the main diagonal.

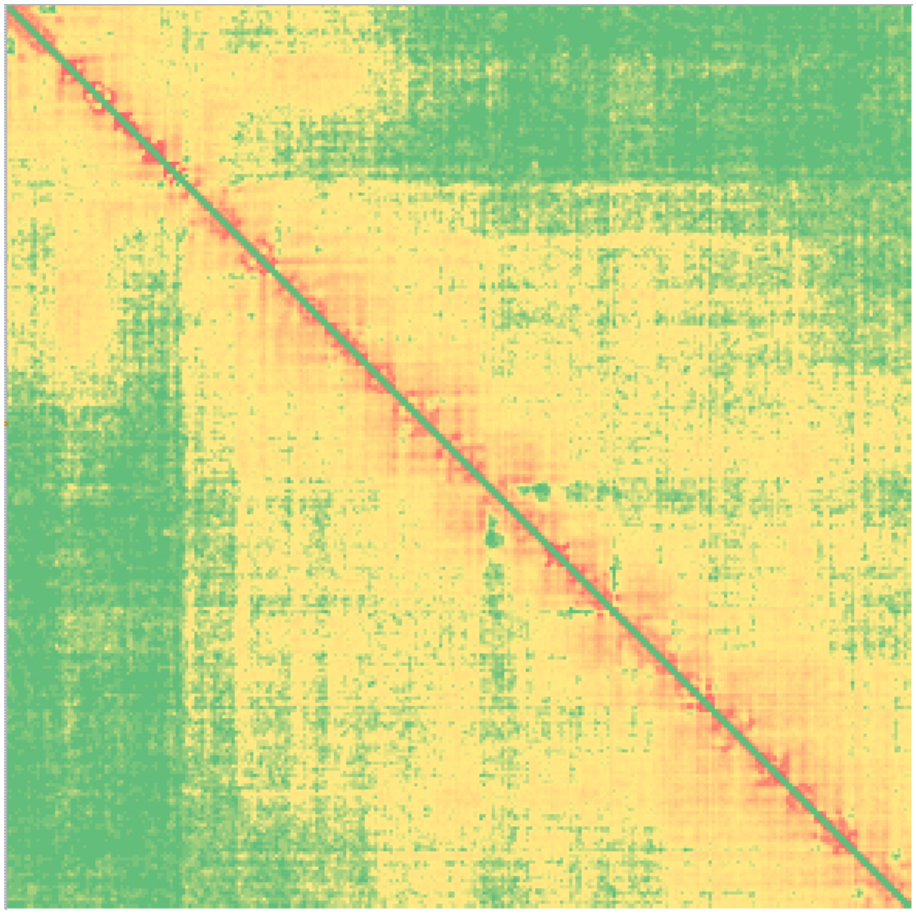


Figure23. Synthetic contact matrix of 100 conformations produced by ChromStruct4 on raw data from Chromosome 1.

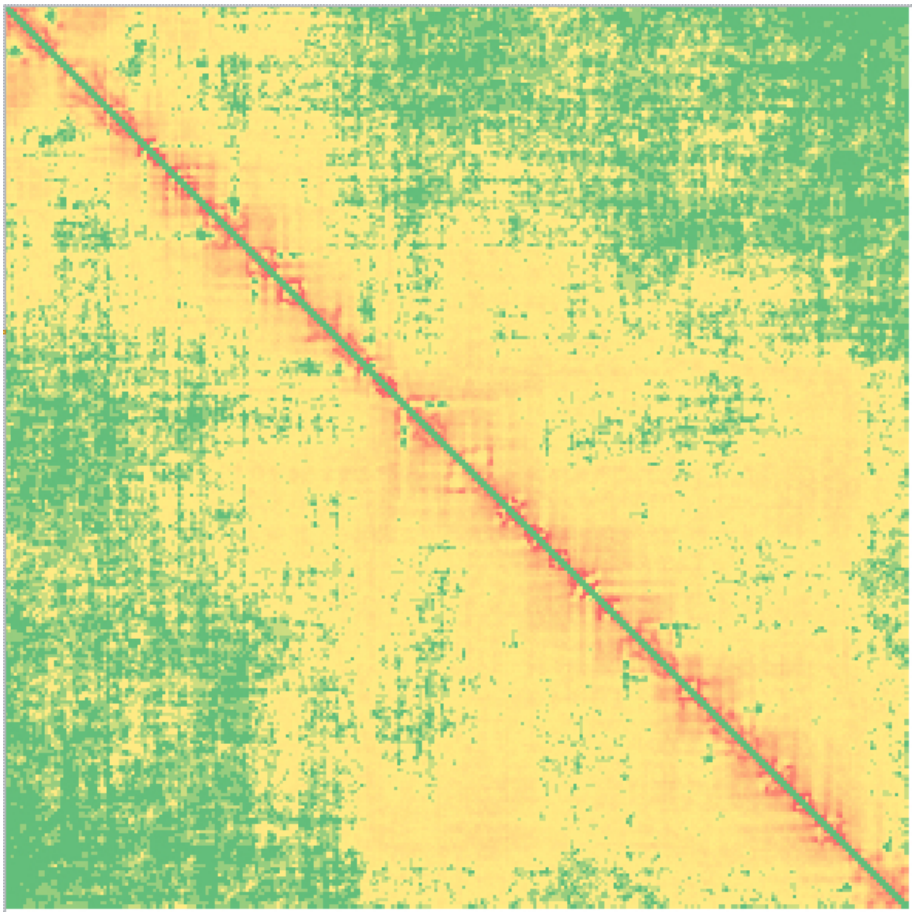


Figure24. Synthetic contact matrix of 100 conformations produced by ChromStruct4 on normalized data from Chromosome 1.

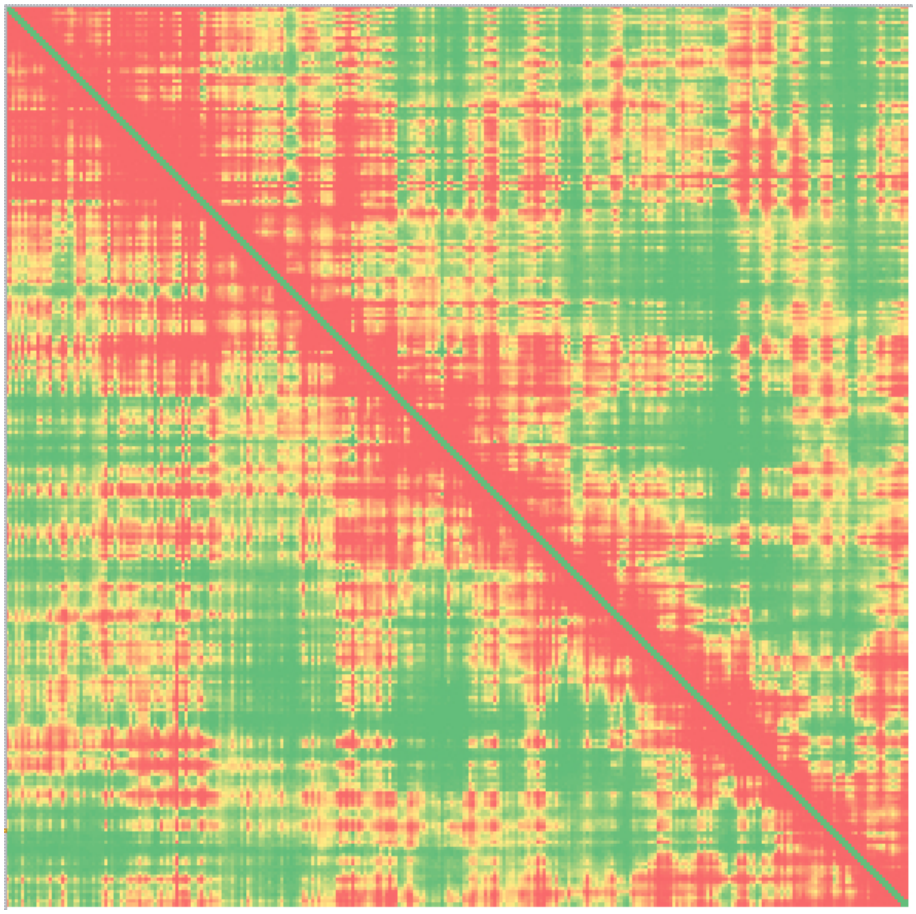


Figure25. Synthetic contact matrix of 100 conformations produced by TADbit on normalized data from Chromosome 1.

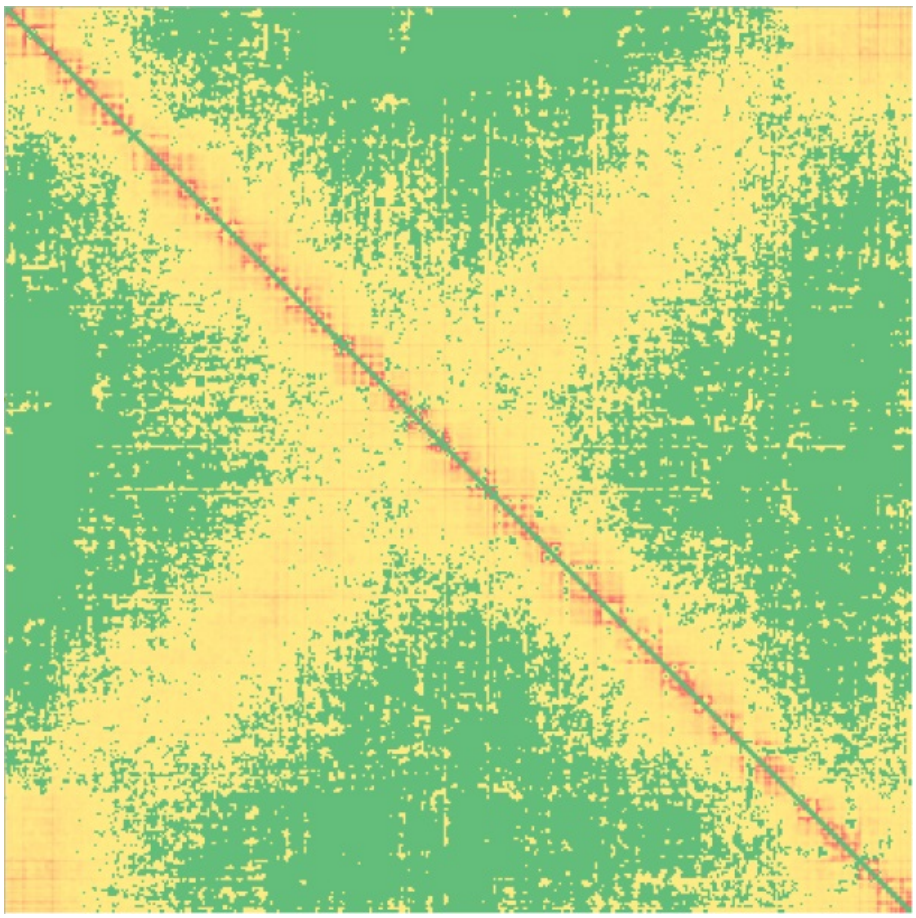


Figure26. Synthetic contact matrix of 100 conformations produced by ChromStruct4 on raw data from Caulobacter.

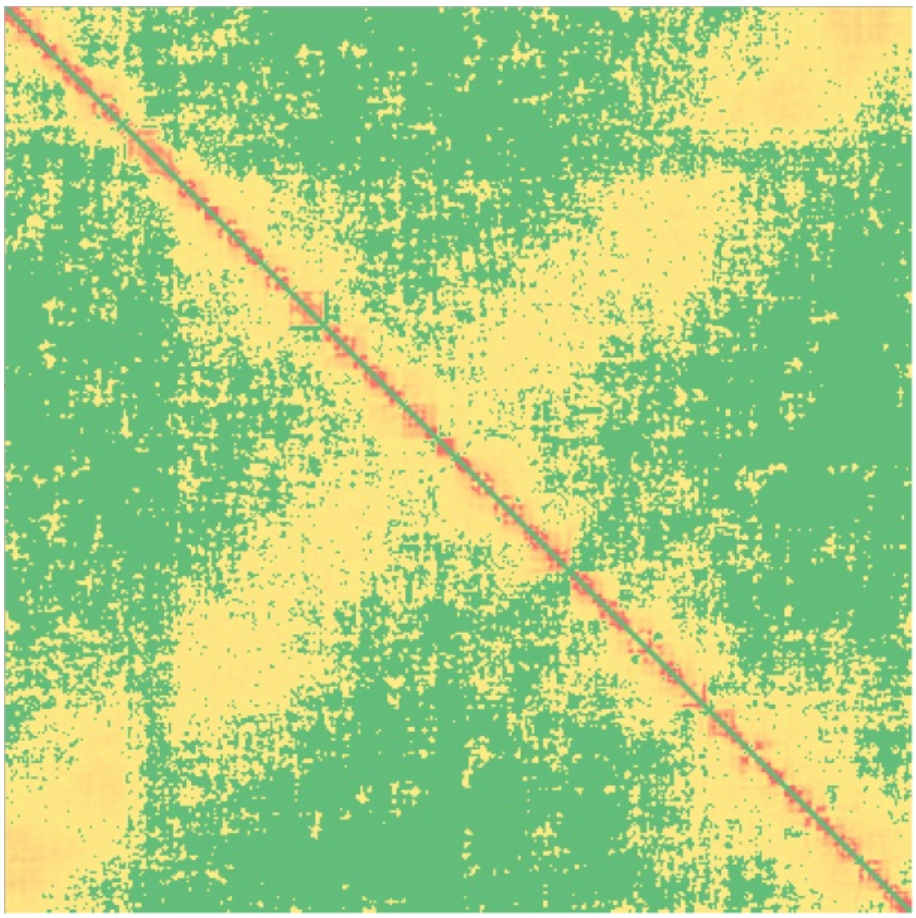


Figure27. Synthetic contact matrix of 100 conformations produced by ChromStruct4 on normalized data from Caulobacter.

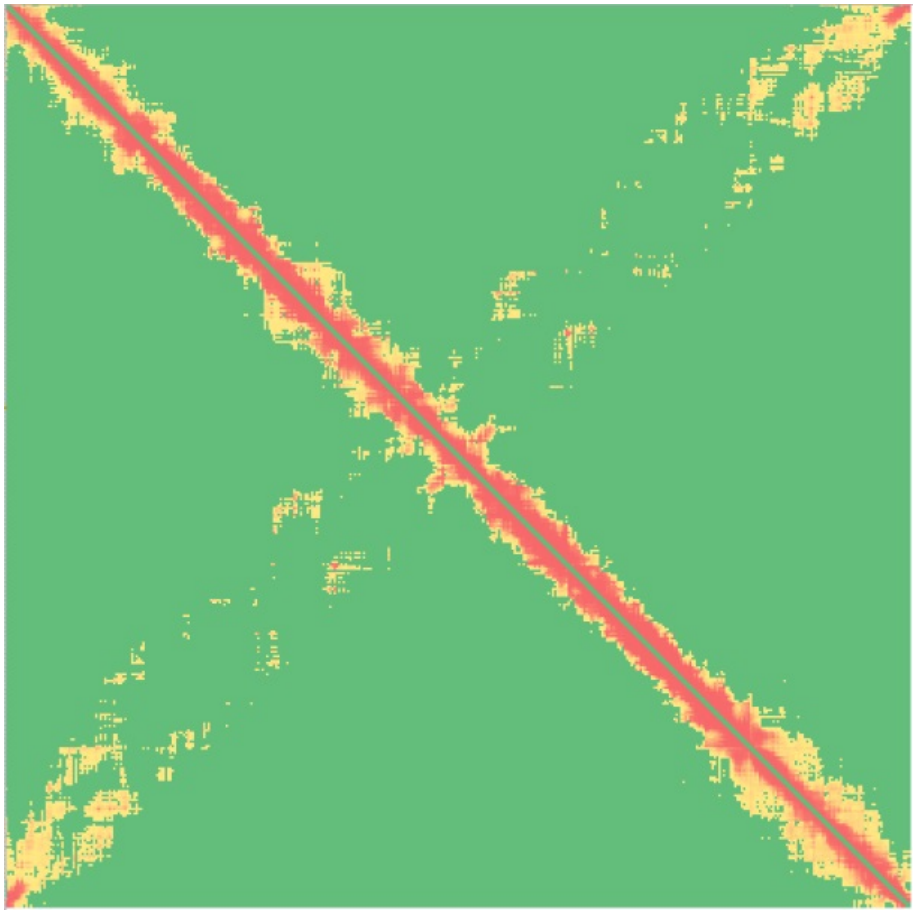


Figure28. Synthetic contact matrix of 100 conformations produced by TADbit on normalized data from Caulobacter.

4 Conclusions

From comparative experiments between ChromStruct4 and TADbit, the following conclusions can be drawn:

- ChromStruct4 is robust against biases, because boxplots Euclidean-distance vs genomic-distance and synthetic matrices, for both Chromosome 1 and Caulobacter, are very similar in raw and normalized data (ICE).
- TADbit and ChromStruct4 produce different conformations, both in structure and sizes. This means that the two approaches focus on different strategies in the chromatin's three-dimensional structure reconstruction. Starting from the same data, they produce different results. The TADbit solutions are also much less variable than those produced by ChromStruct4.

References

1. Baú, D. *et.al.* (2012): Genome structure determination via 3C -based data integration by the Integrative Modeling Platform, Methods, 58(3): 300306
2. Lieberman-Aiden, E. *et.al.* (2009): Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome, Science 326: 289-293.
3. Imakaev, M. *et.al.* (2012): Iterative correction of HI-C data reveals hallmarks of chromosome organization, Nature Methods 9(10): 999-1003.
4. Dixon, J.R. *et.al.* (2012): Topological domains in mammalian genomes identified by analysis ofchromatin interactions, Nature 485: 376-380.
5. Caudai, C. *et.al.* (2015): Inferring 3D chromatin structure using a multiscale approach based on quaternions, BMC Bioinformatics, 16: 234.
6. Caudai, C. *et.al.* (2015): A Statistical Approach to Infer 3D Chromatin Structure, Mathematical Models in Biology, Springer, 161-171.