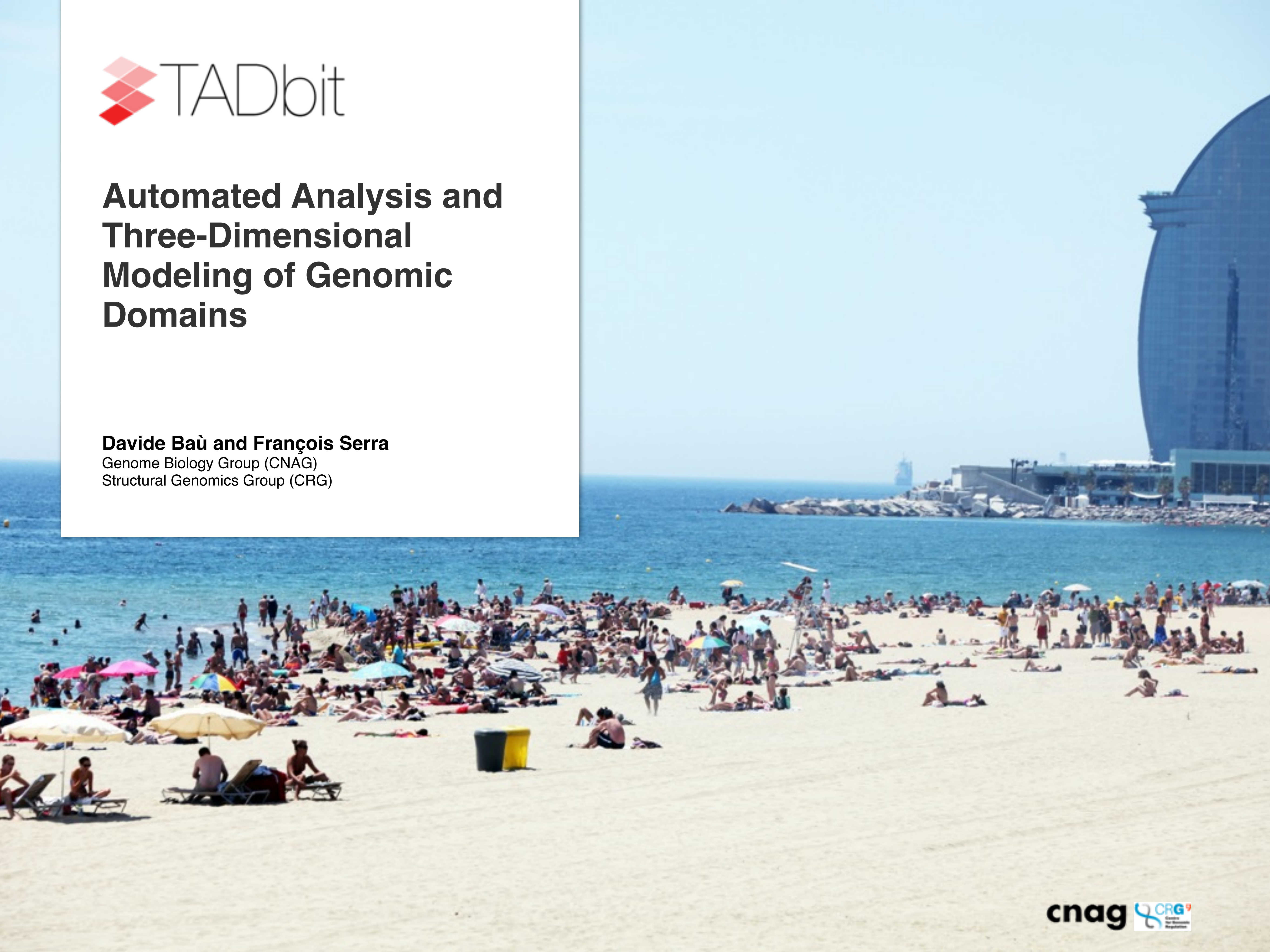


TADbit

Automated Analysis and Three-Dimensional Modeling of Genomic Domains

Davide Baù and François Serra
Genome Biology Group (CNAG)
Structural Genomics Group (CRG)

cnag CRG

Tutorial outline

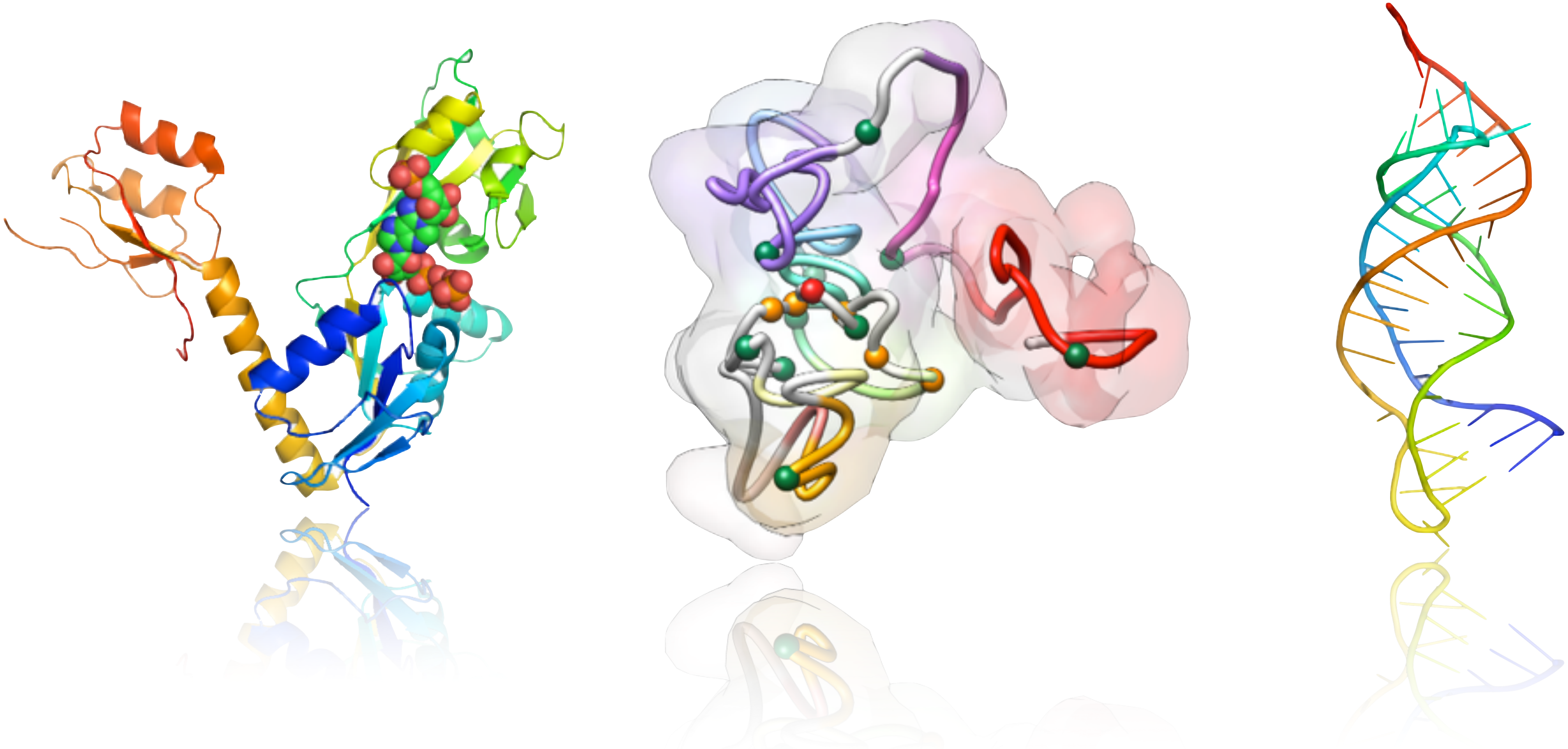
Theory Practice

Morning
- Chromatin structure and Hi-C data
- The Integrative Modeling Platform applied to chromatin
- TADbit introduction and installation

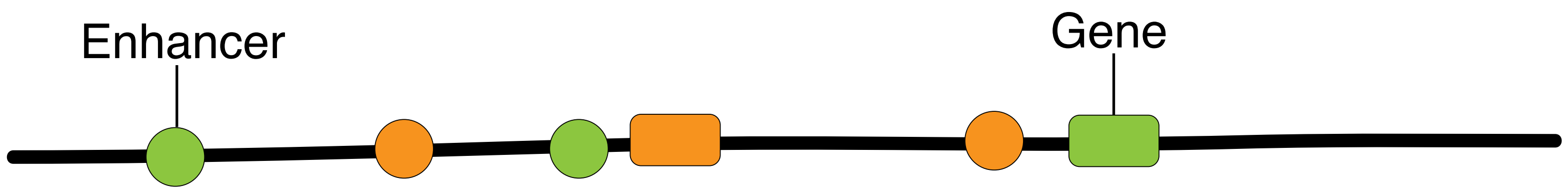
Afternoon
- Topologically Associated Domains detection and analysis
- 3D modeling of real Hi-C data
- Analysis of the results

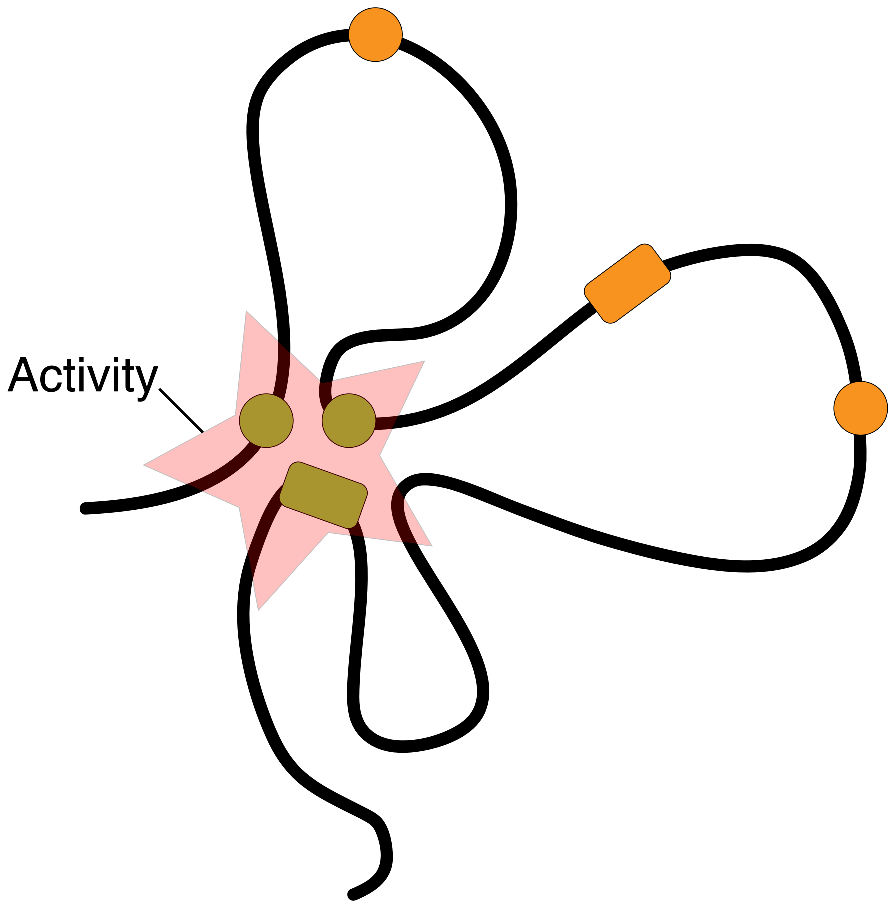
Structural Genomics Group

http://www.marciuslab.org

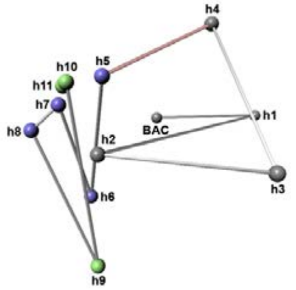


The genome is not linearly organized

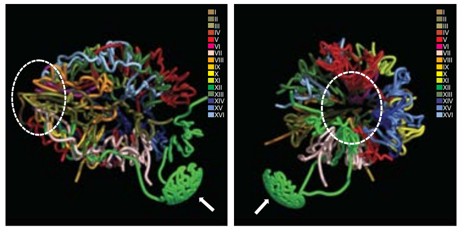




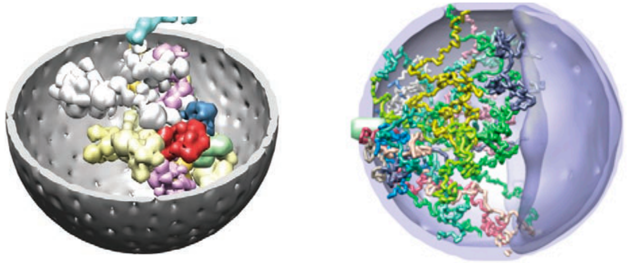
3D modeling of genomic domains: other methods



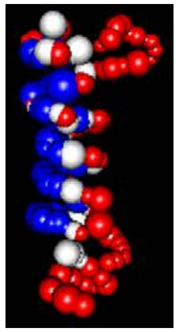
Jhunjhunwala (2008) Cell



Duan (2010) Nature

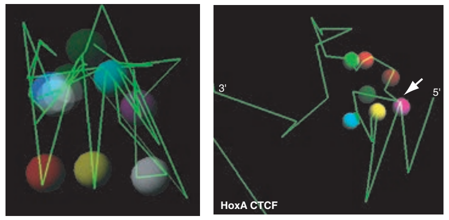


Kalhor (2011) Nature Biotechnology
Tjong (2012) Genome Research

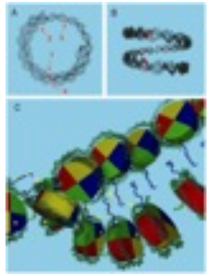


Hu (2013) PLoS Computational Biology

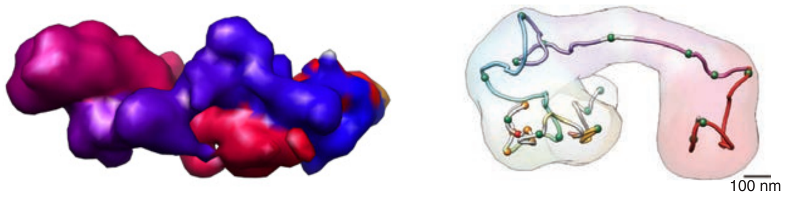
2008 2014



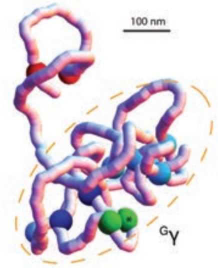
Fraser (2009) Genome Biology
Ferraiuolo (2010) Nucleic Acids Research



Asbury (2010) BMC Bioinformatics



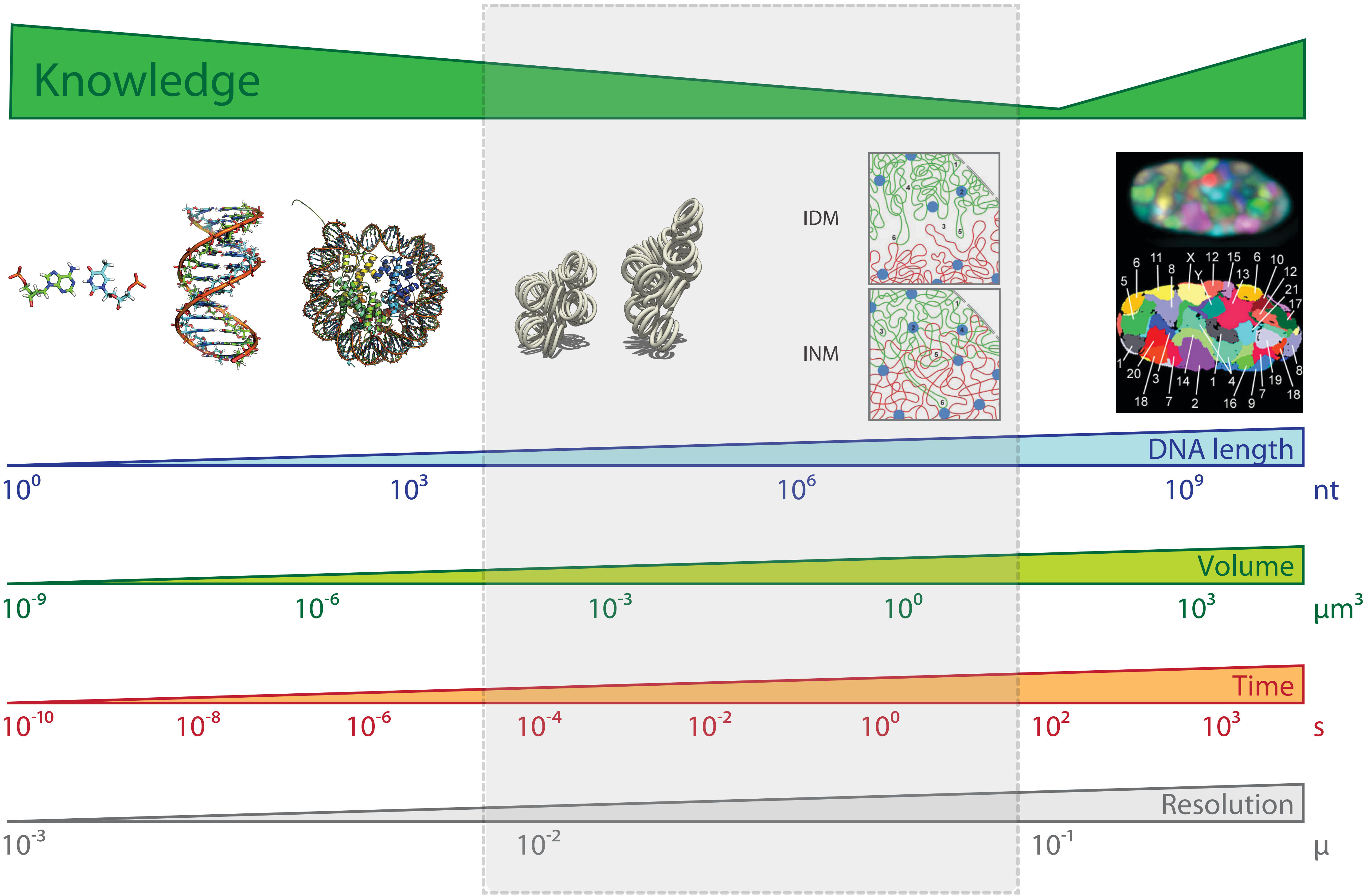
Baù (2011) Nature Structural & Molecular Biology
Umbarger (2011) Molecular Cell



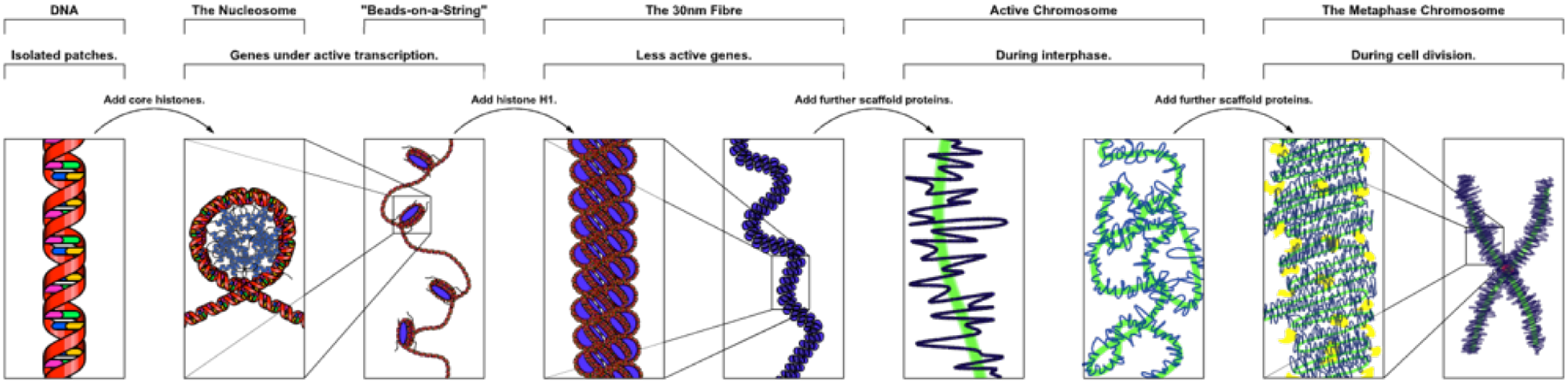
Junier (2012) Nucleic Acids Research

Resolution gap

Marti-Renom, M. A. & Mirny, L. A. PLoS Comput Biol 7, e1002125 (2011)

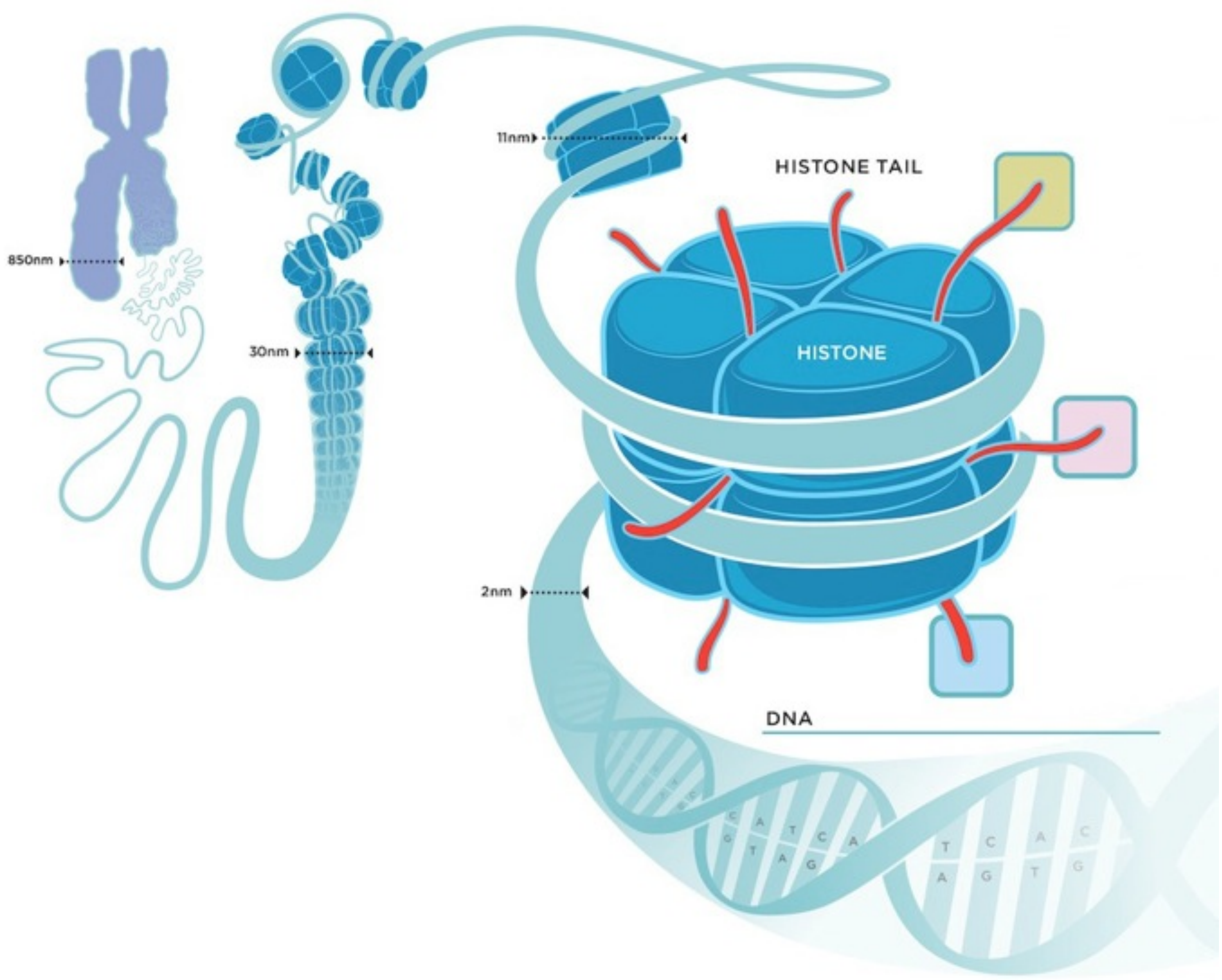


Chromatin structure: different levels of organization



The nuclear organization of DNA

Chromosome **Chromatin fibre** **Nucleosome**



Adapted from Richard E. Ballermann, 2012

Chromatin definition

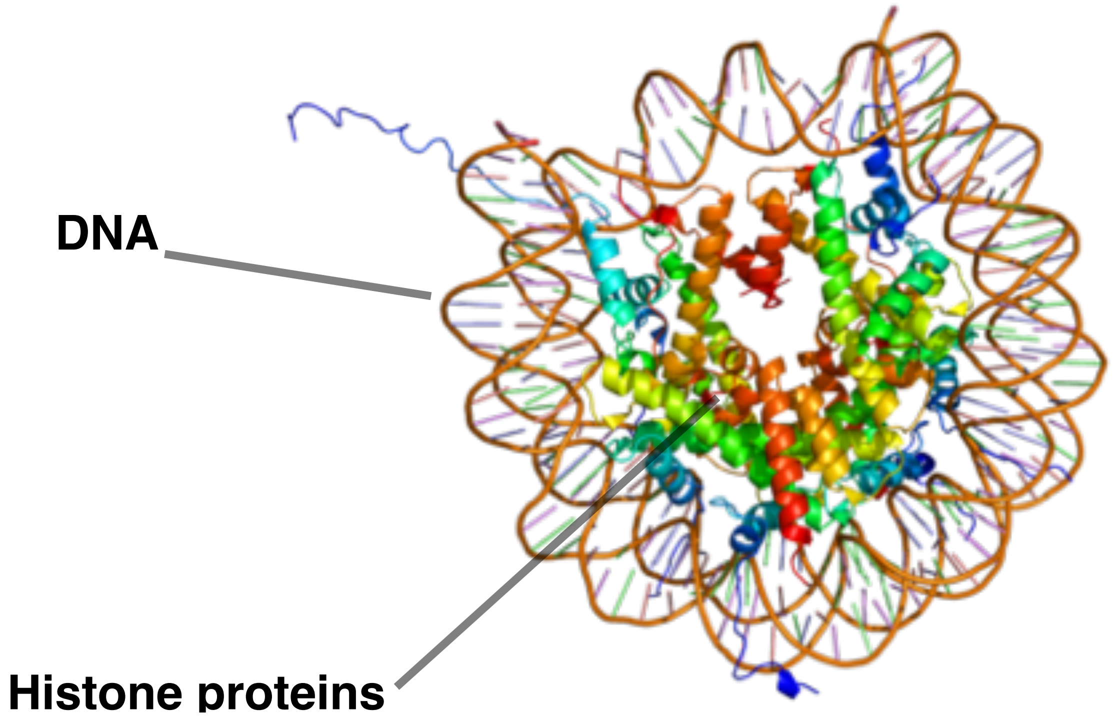
Chromatin is composed of DNA complexed with histones and other proteins

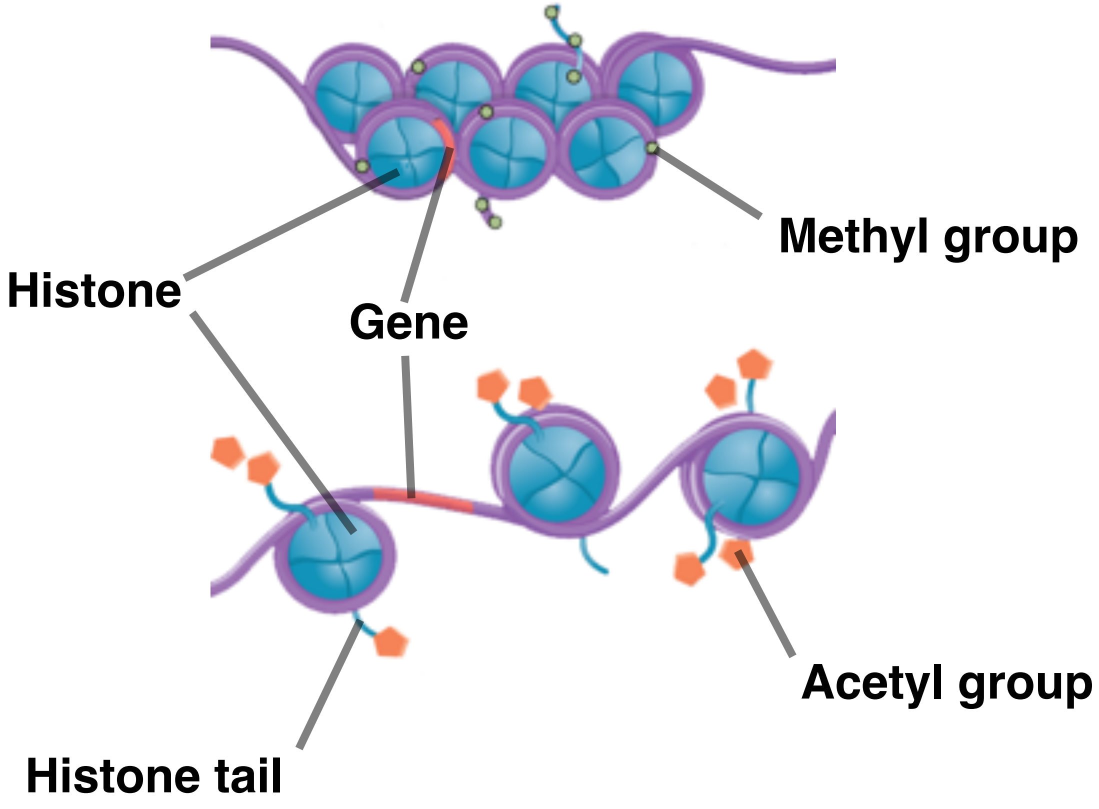
Chromatin formation enables the genome to be hierarchically packaged or condensed so that it can fit inside the nuclear space

The compaction allows to modulate gene transcription, DNA repair, recombination, and replication

Chromatin structure is considered highly dynamic

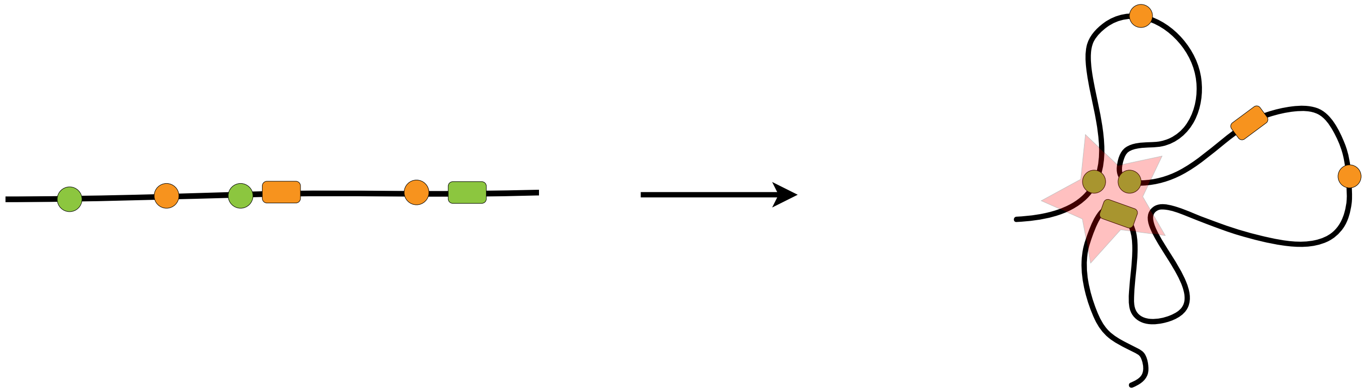
The nucleosome



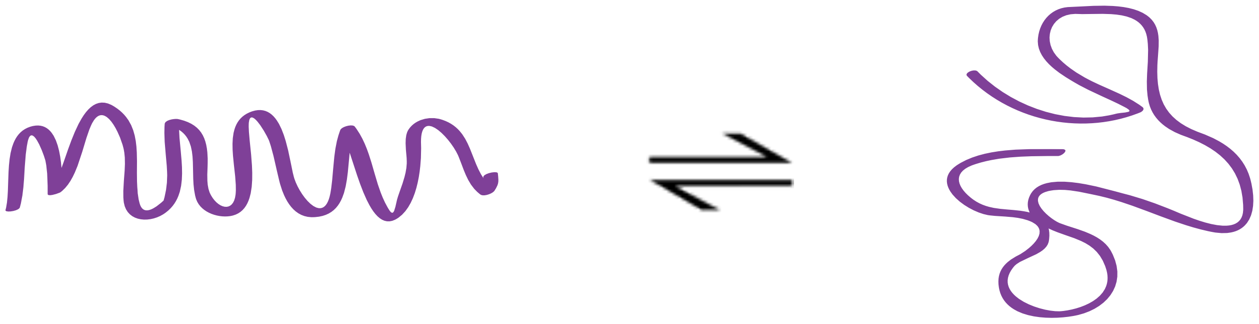


The role of chromatin structure

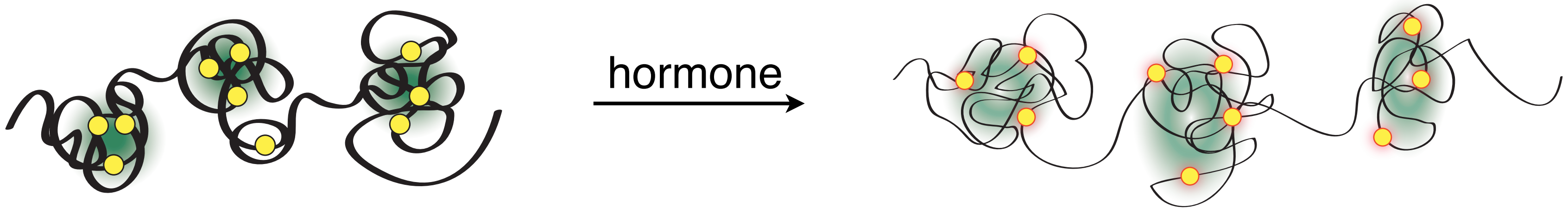
Activity



Organization

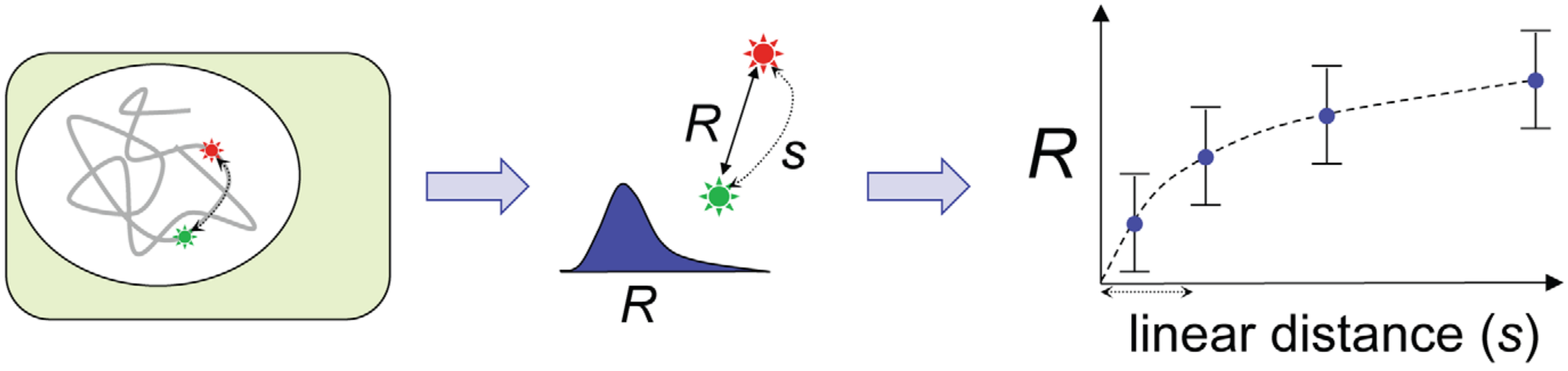


Processes

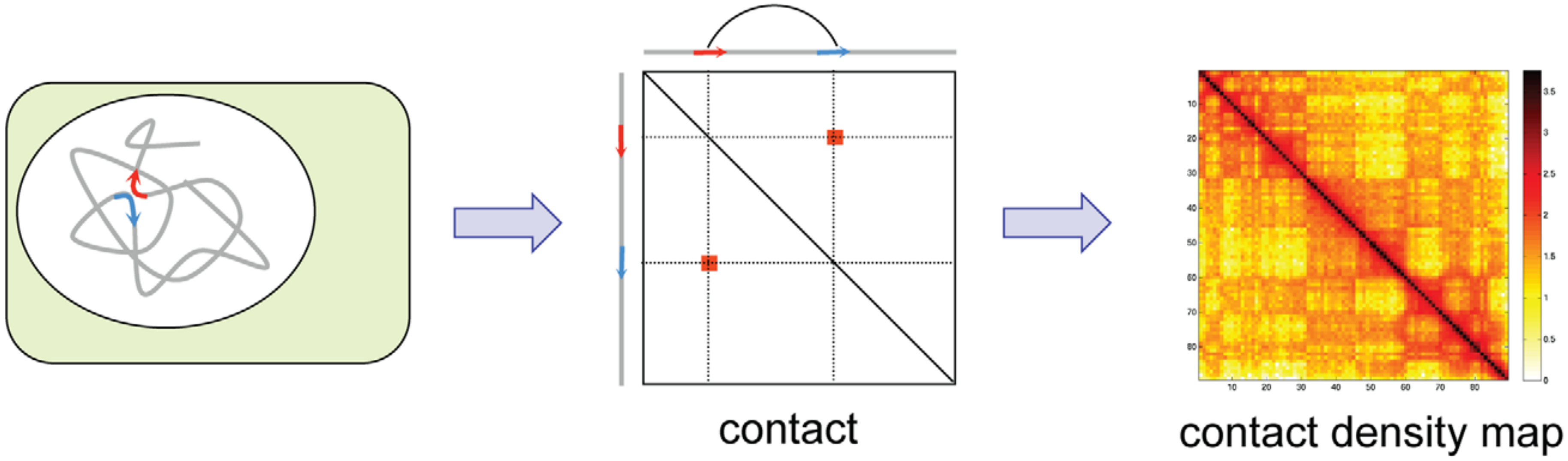


Data used for chromatin structure determination

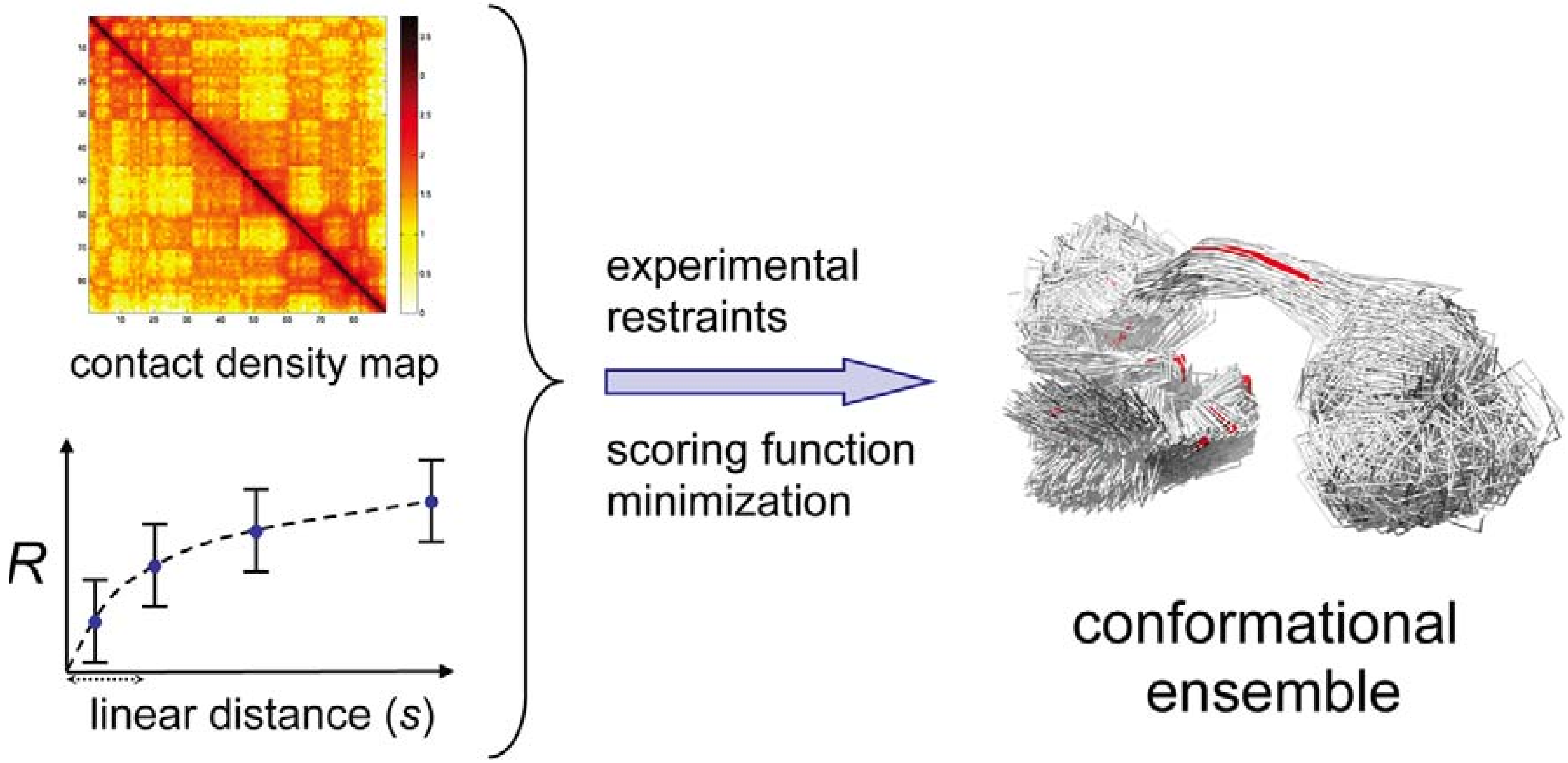
Light microscopy (FISH)



Cell/molecular biology (3C-based methods)



Restrain based modeling (IMP)



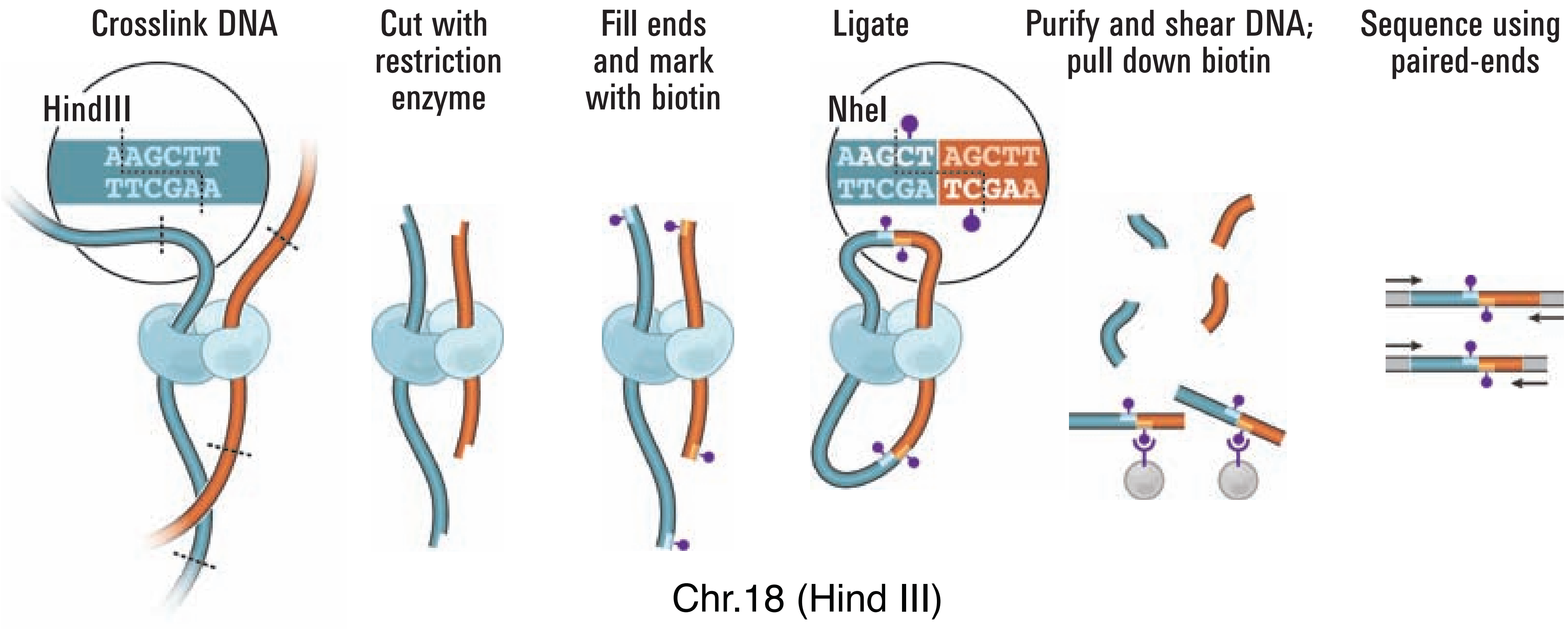
experimental restraints

scoring function minimization

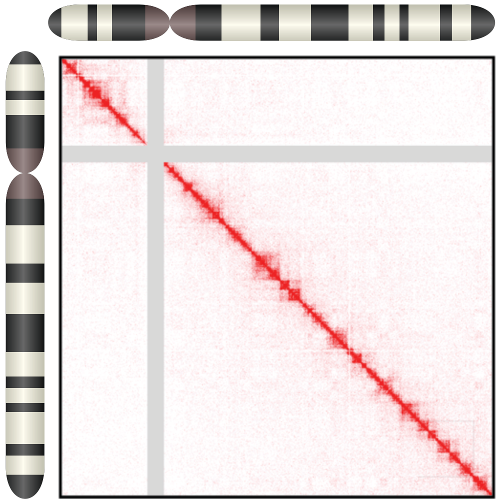
conformational ensemble

Hi-C technology

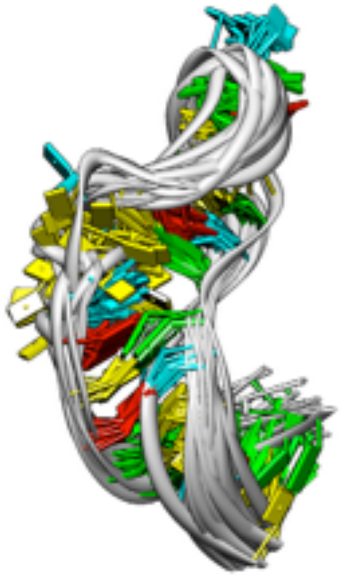
Lieberman-Aiden, E. et al. Science 326, 289–293 (2009).
http://3dg.umassmed.edu

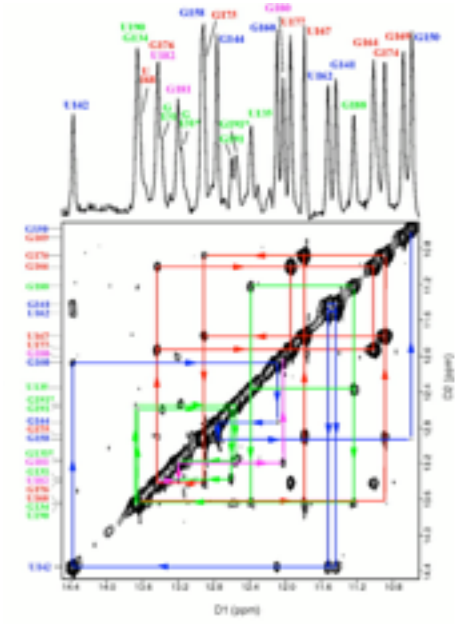


Chr.18 (Hind III)

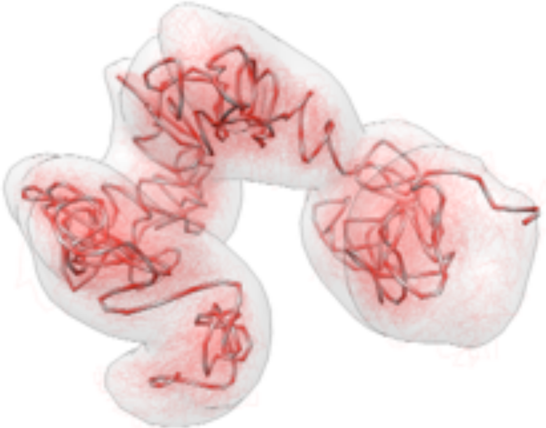


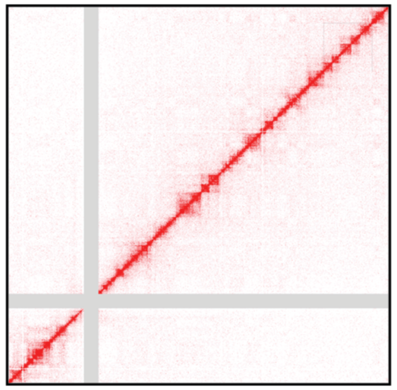
Structure determination using Hi-C data





Biomolecular structure determination
2D-NOESY data



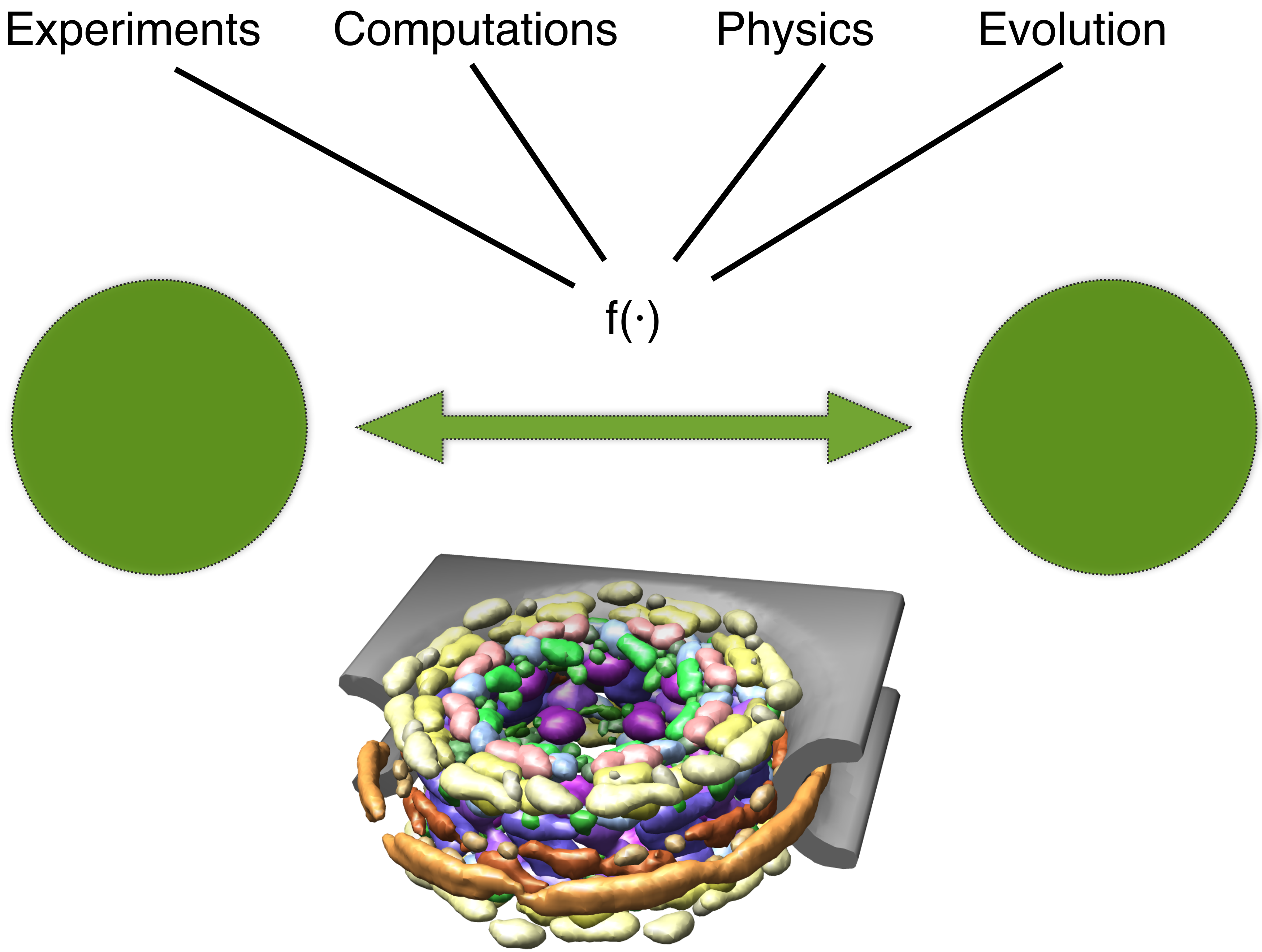


Chromosome structure determination
3C-based data

The Integrative Modeling Platform framework

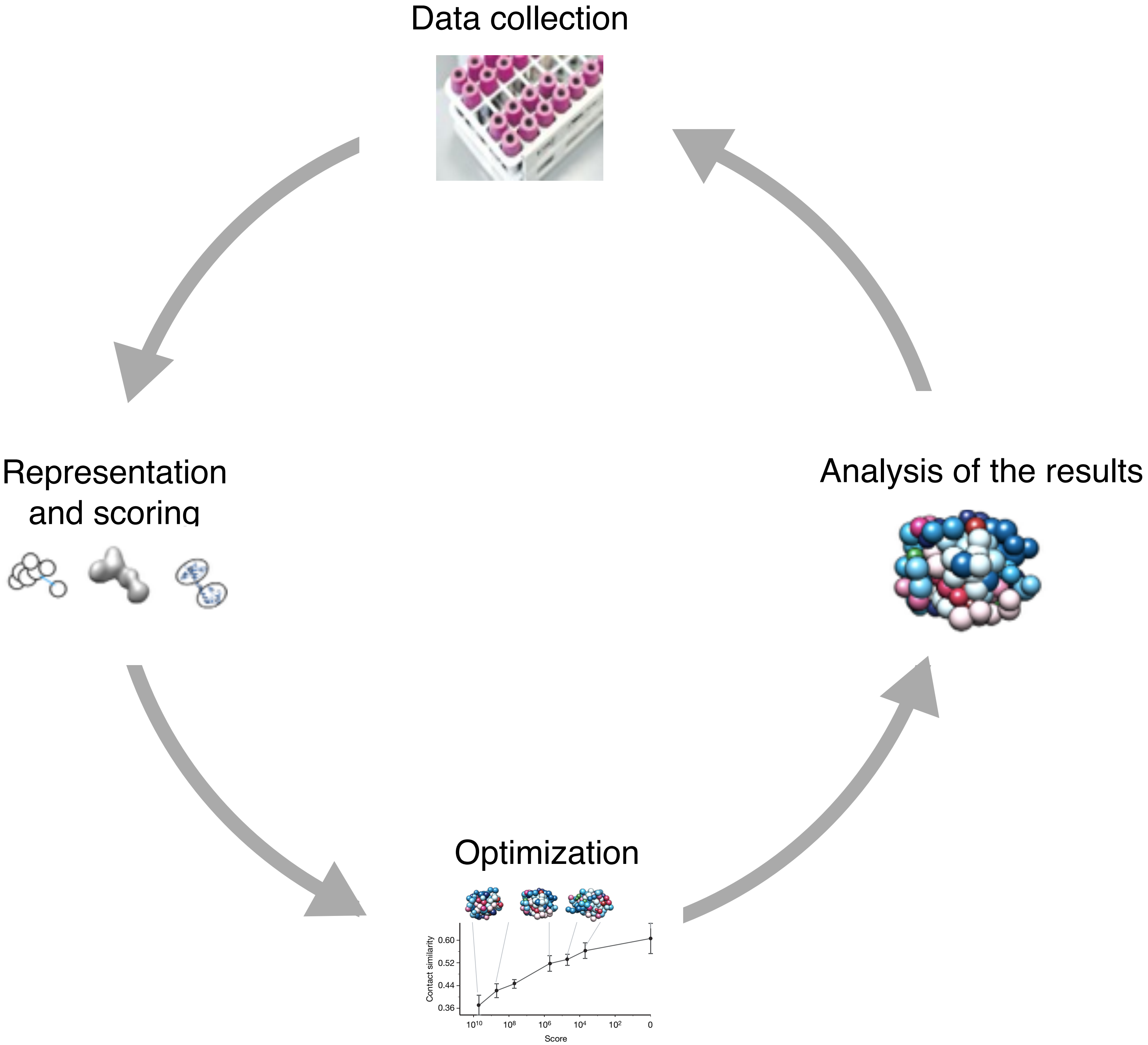
http://www.integrativemodeling.org

Russel, D. et al. PLOS Biology 10, e1001244 (2012).



From Alber, F. et al. Nature 450, 695–701 (2007).

The four stages of integrative modeling



Data collection

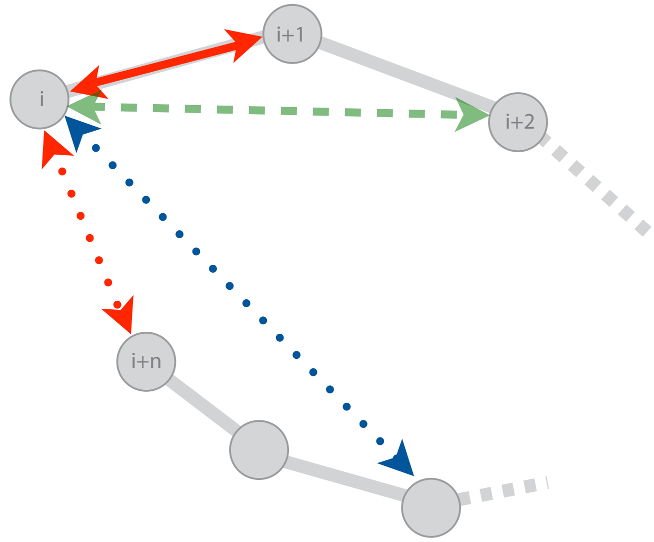
Representation and scoring

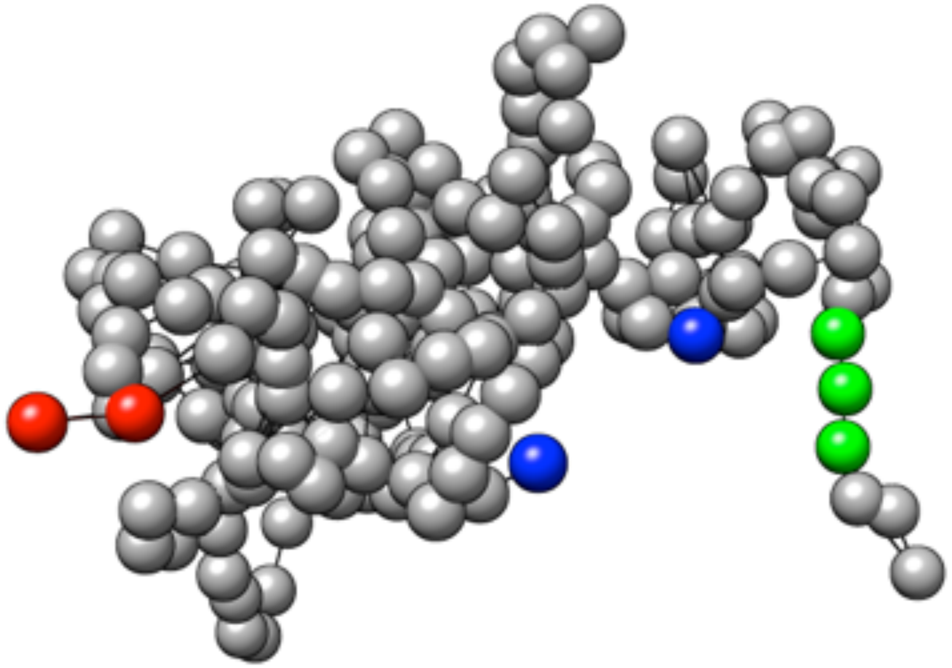
Optimization

Analysis of the results

Model representation and scoring

Constituent parts of the molecule

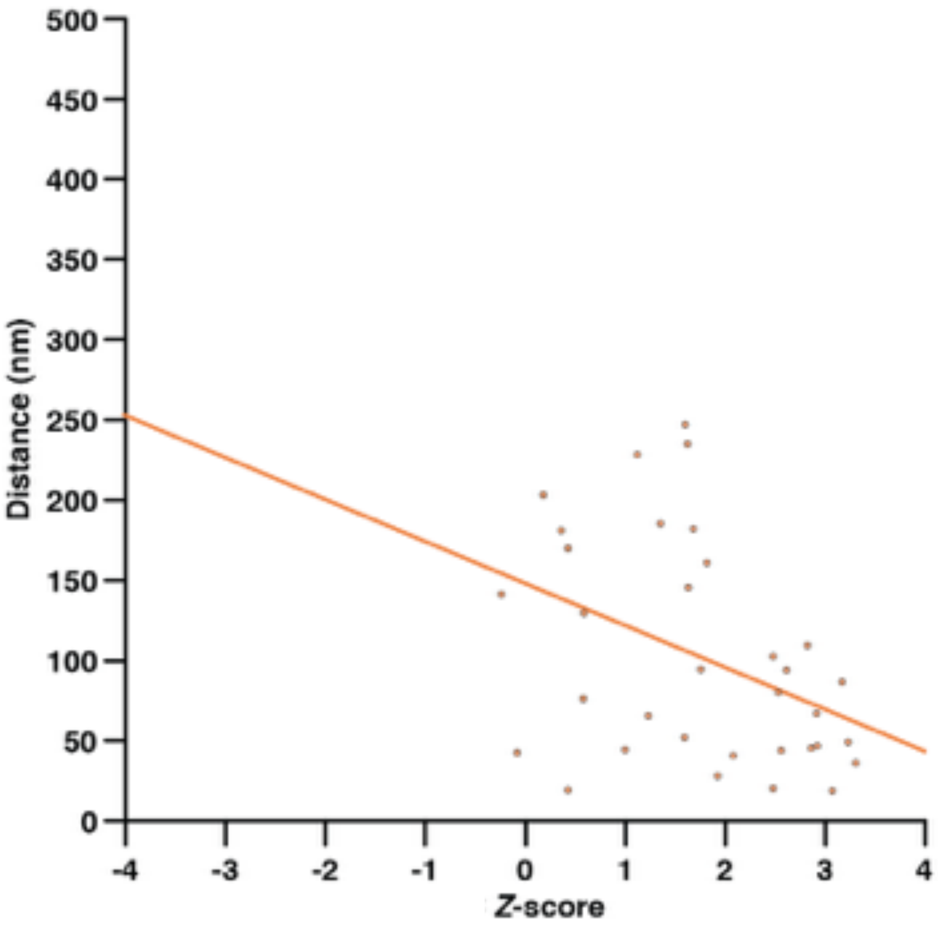




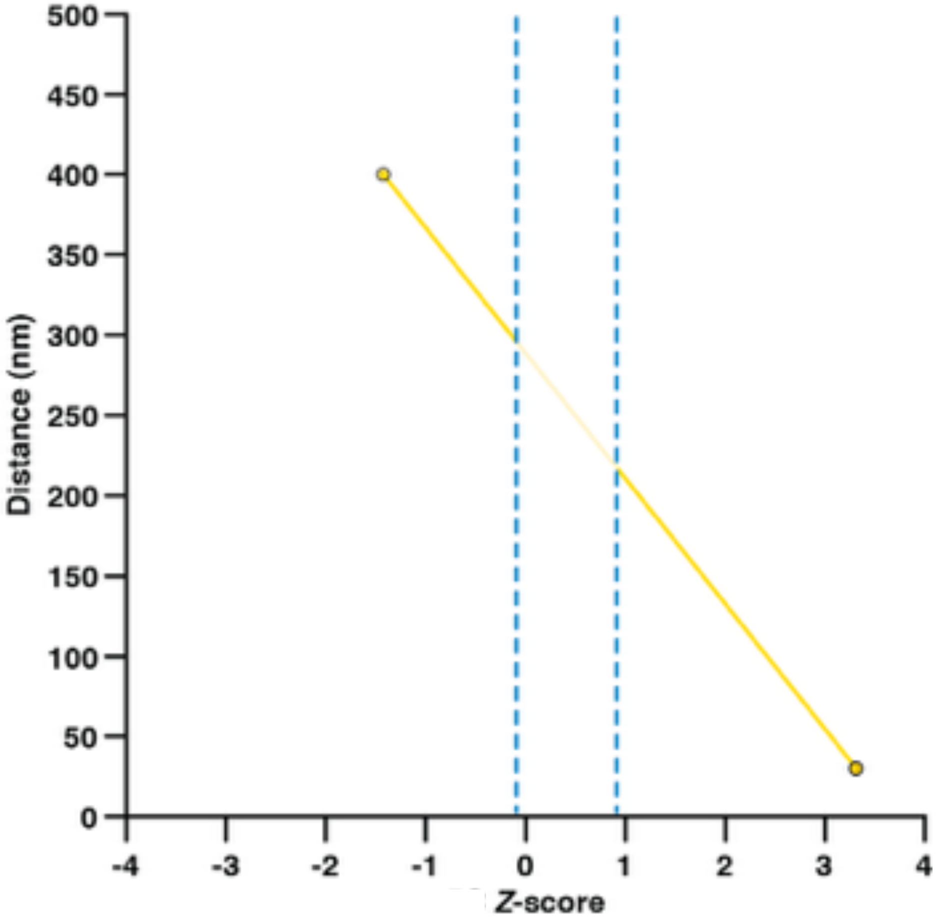
$d < d_0$ $d = d_0$ $d > d_0$

3C-like data to spatial distances

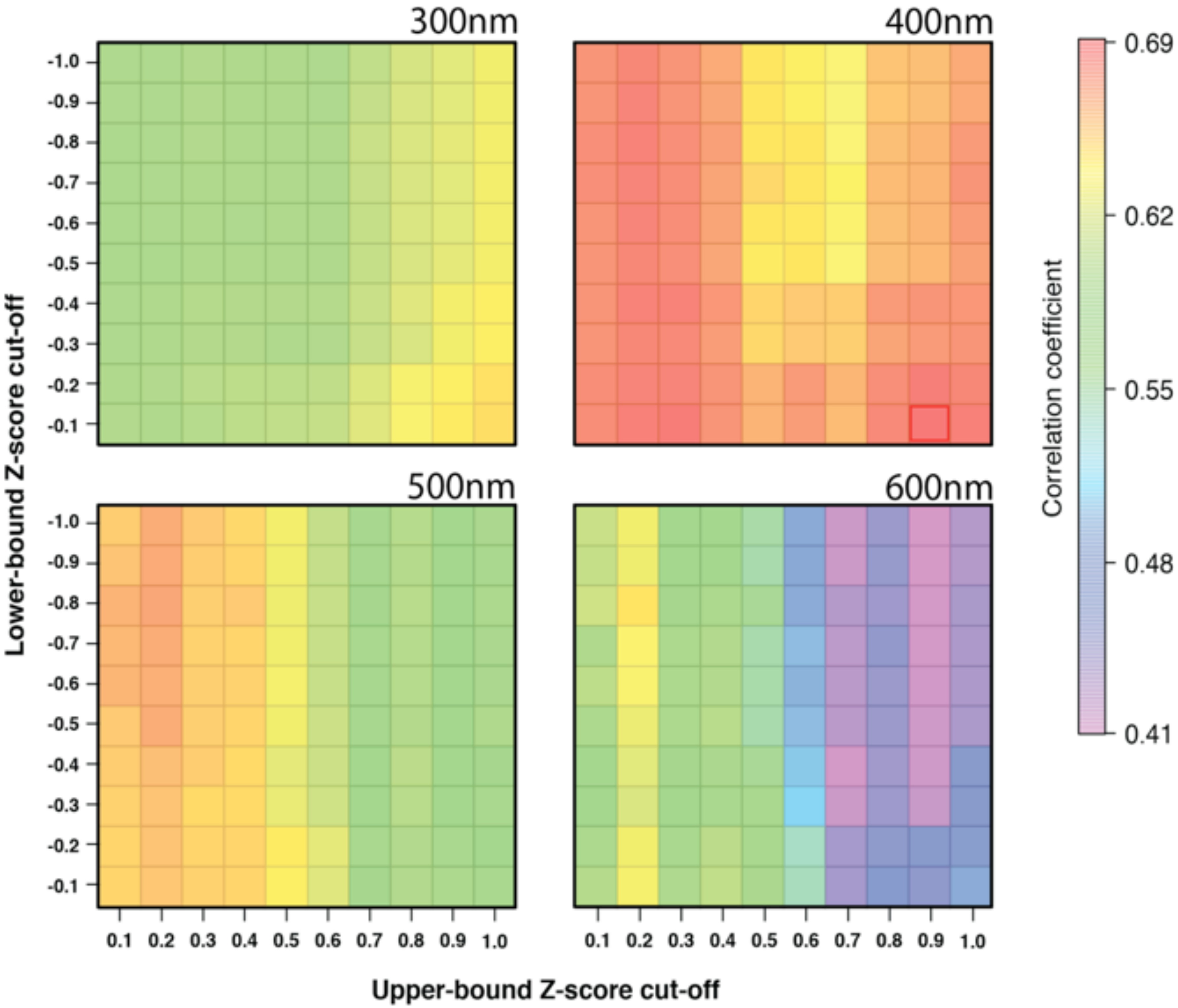
Neighbor fragments



Non-Neighbor fragments



Parameter optimization

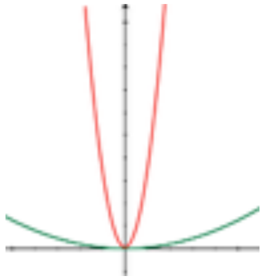


Model representation and scoring

Constituent parts of the molecule

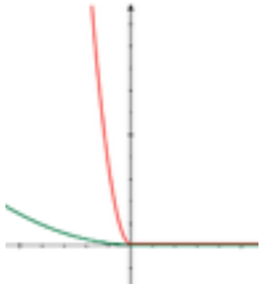
Harmonic

$$H_{i,j} = k\left(d_{i,j} - d^0_{i,j}\right)^2$$



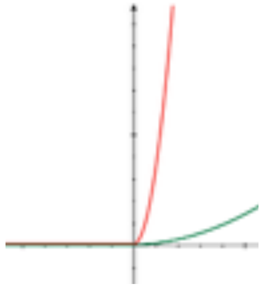
Harmonic Lower Bound

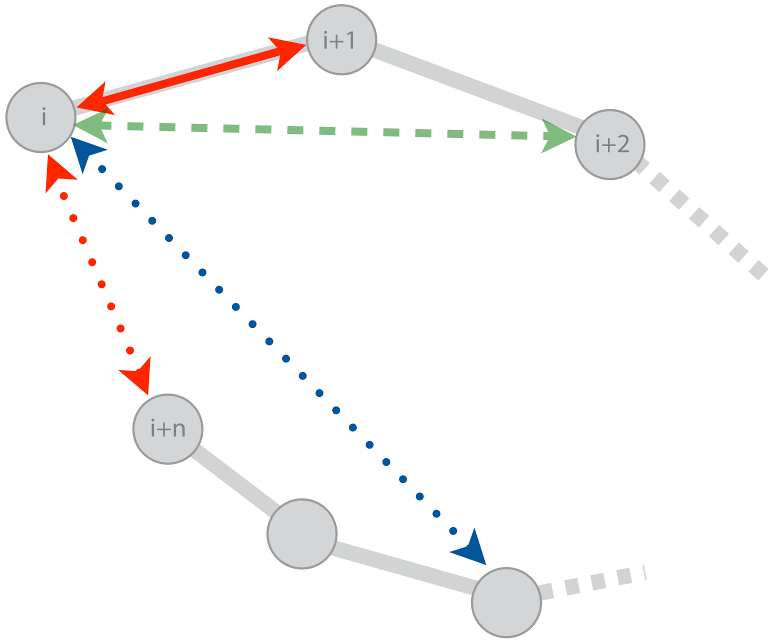
$$\begin{cases} if\ \ d_{i,j} \le d^0_{i,j}; & lbH_{i,j} = k\left(d_{i,j} - d^0_{i,j}\right)^2 \\ if\ \ d_{i,j} > d^0_{i,j}; & lbH_{i,j} = 0 \end{cases}$$



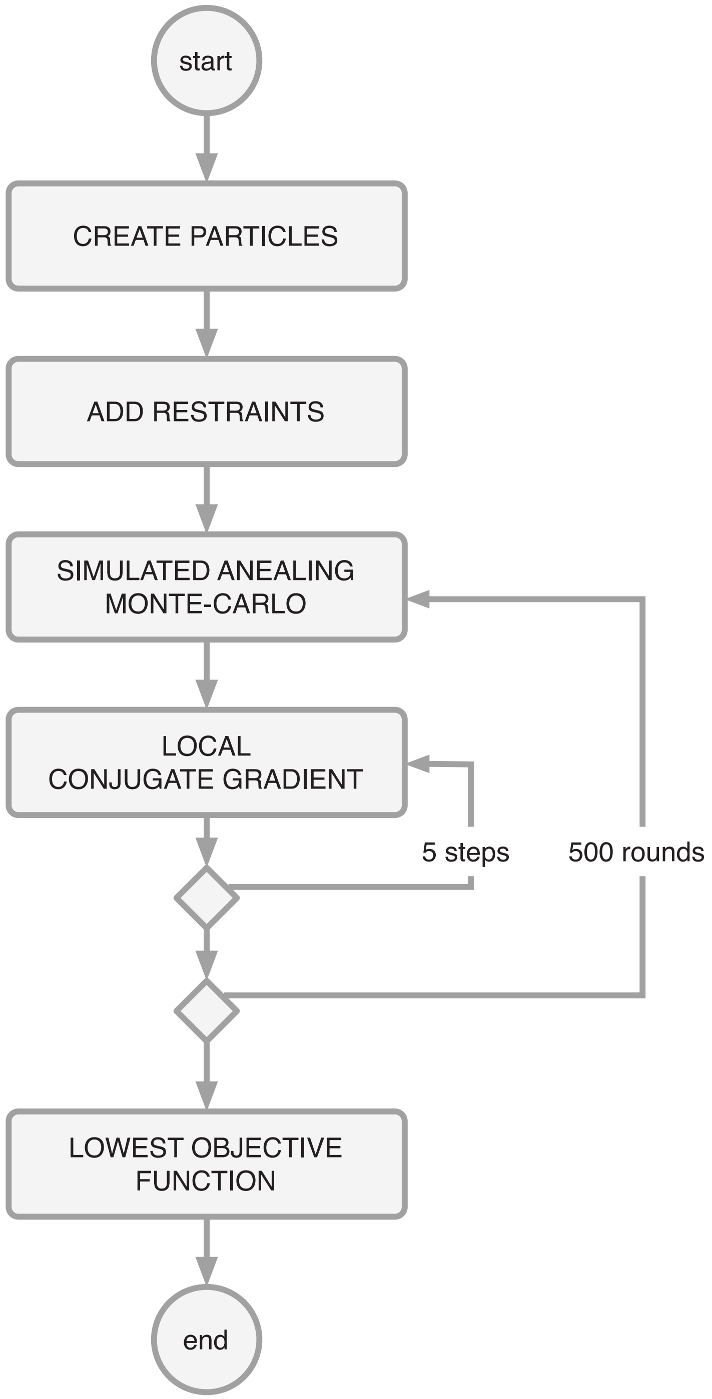
Harmonic Upper Bound

$$\begin{cases} if\ \ d_{i,j} \ge d^0_{i,j}; & ubH_{i,j} = k\left(d_{i,j} - d^0_{i,j}\right)^2 \\ if\ \ d_{i,j} < d^0_{i,j}; & ubH_{i,j} = 0 \end{cases}$$





Optimization of the scoring function



start

CREATE PARTICLES

ADD RESTRAINTS

SIMULATED ANEALING
MONTE-CARLO

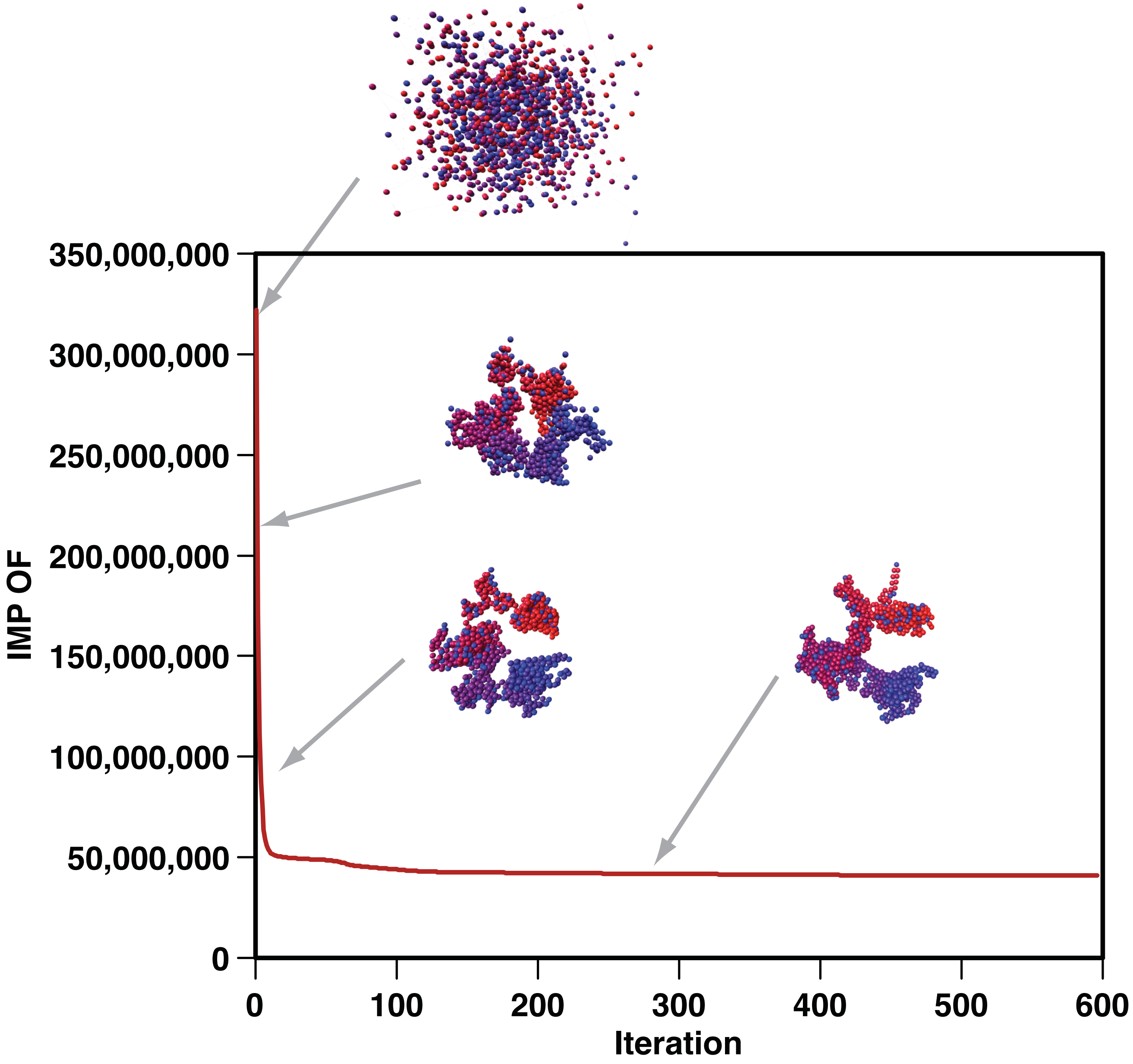
LOCAL
CONJUGATE GRADIENT

5 steps

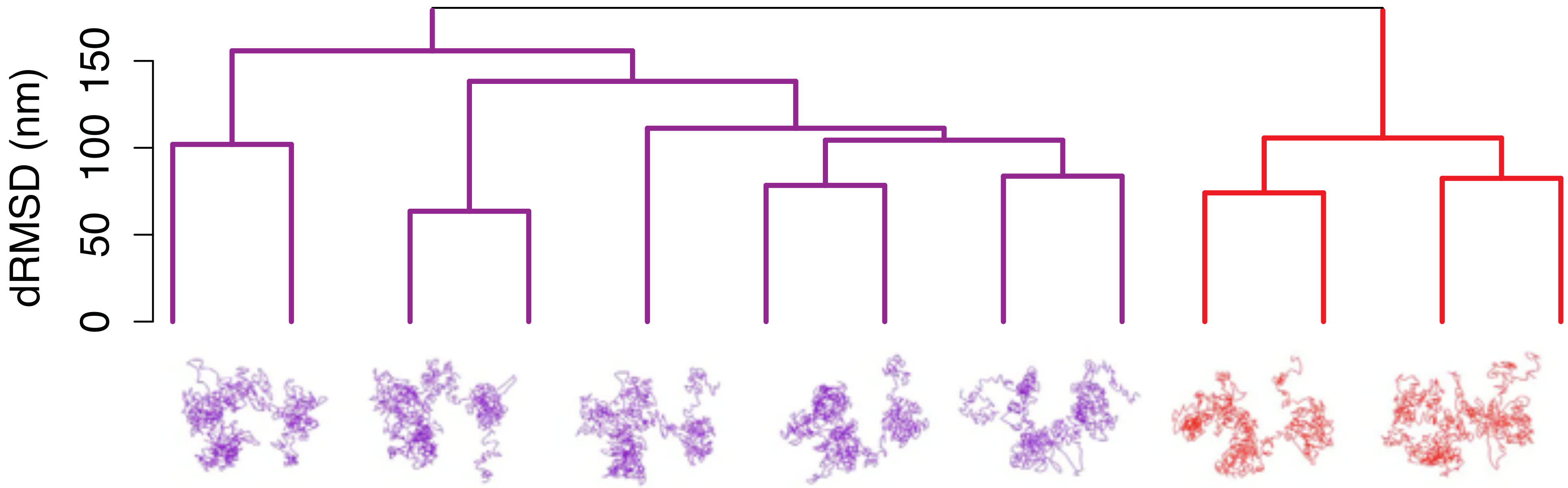
500 rounds

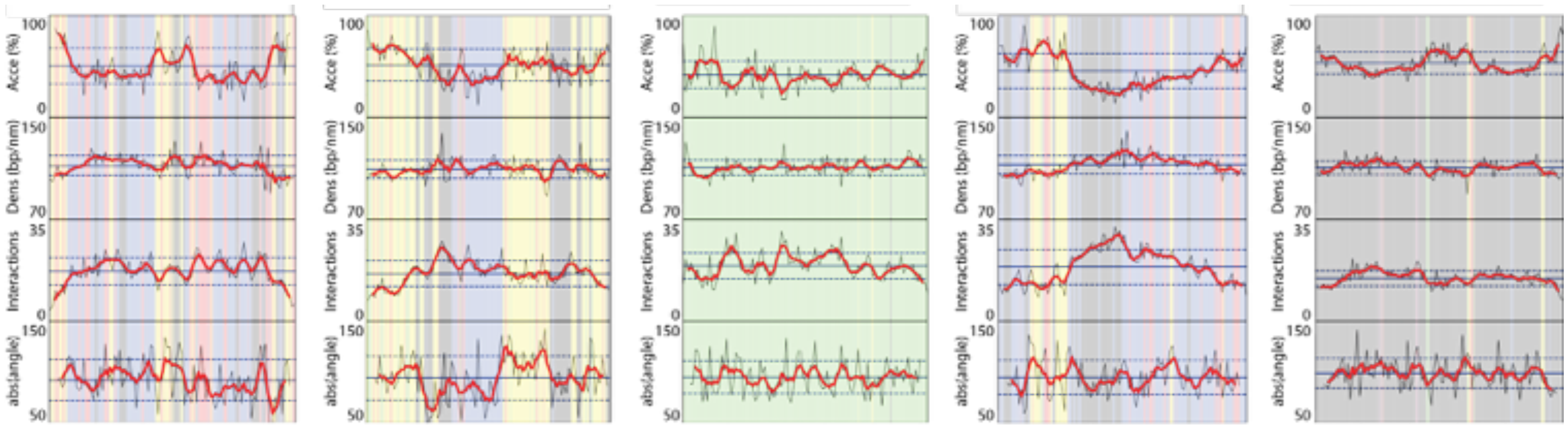
LOWEST OBJECTIVE
FUNCTION

end



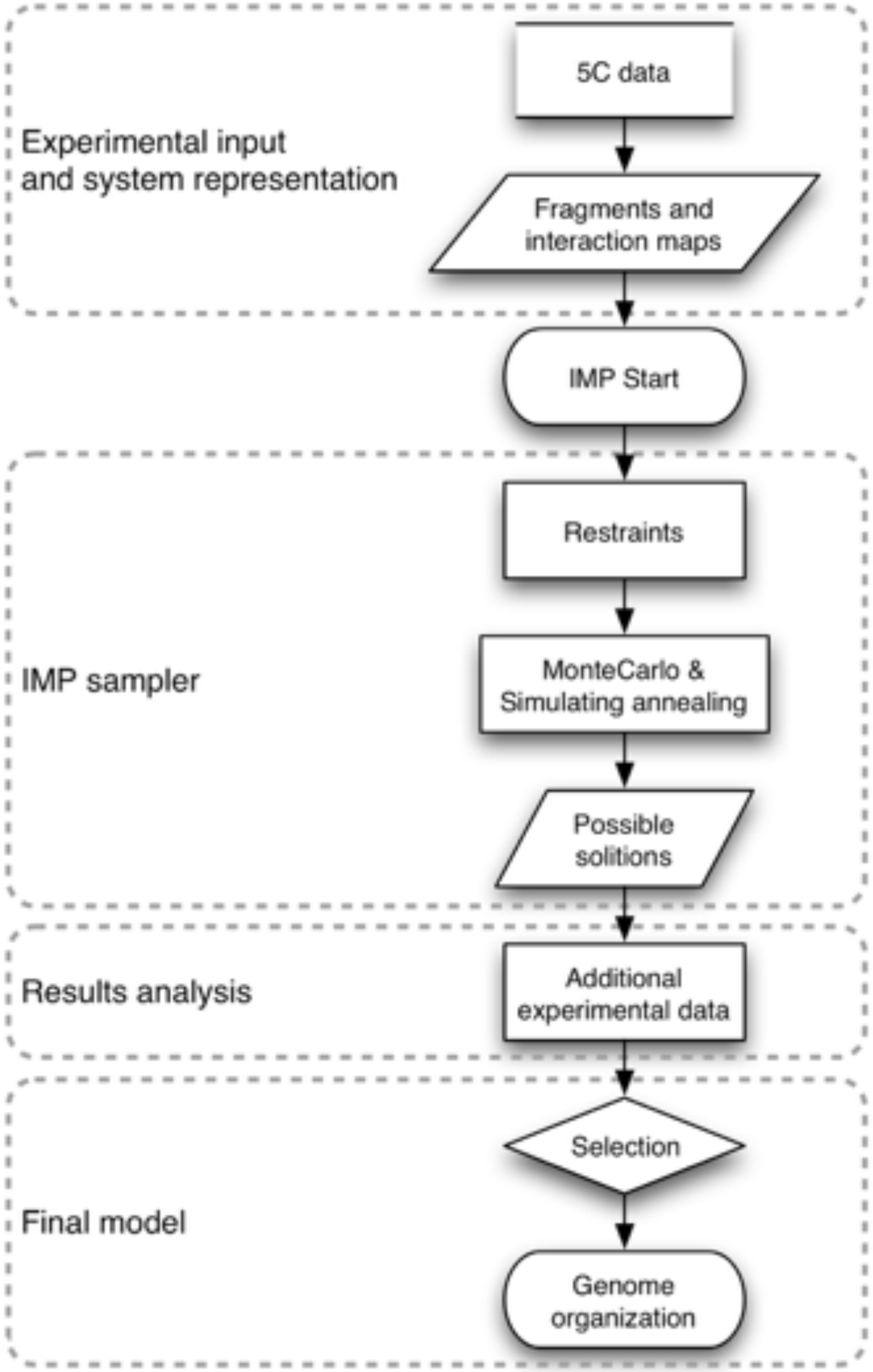
Clustering and structural features





Integrative Modeling

http://www.integrativemodeling.org



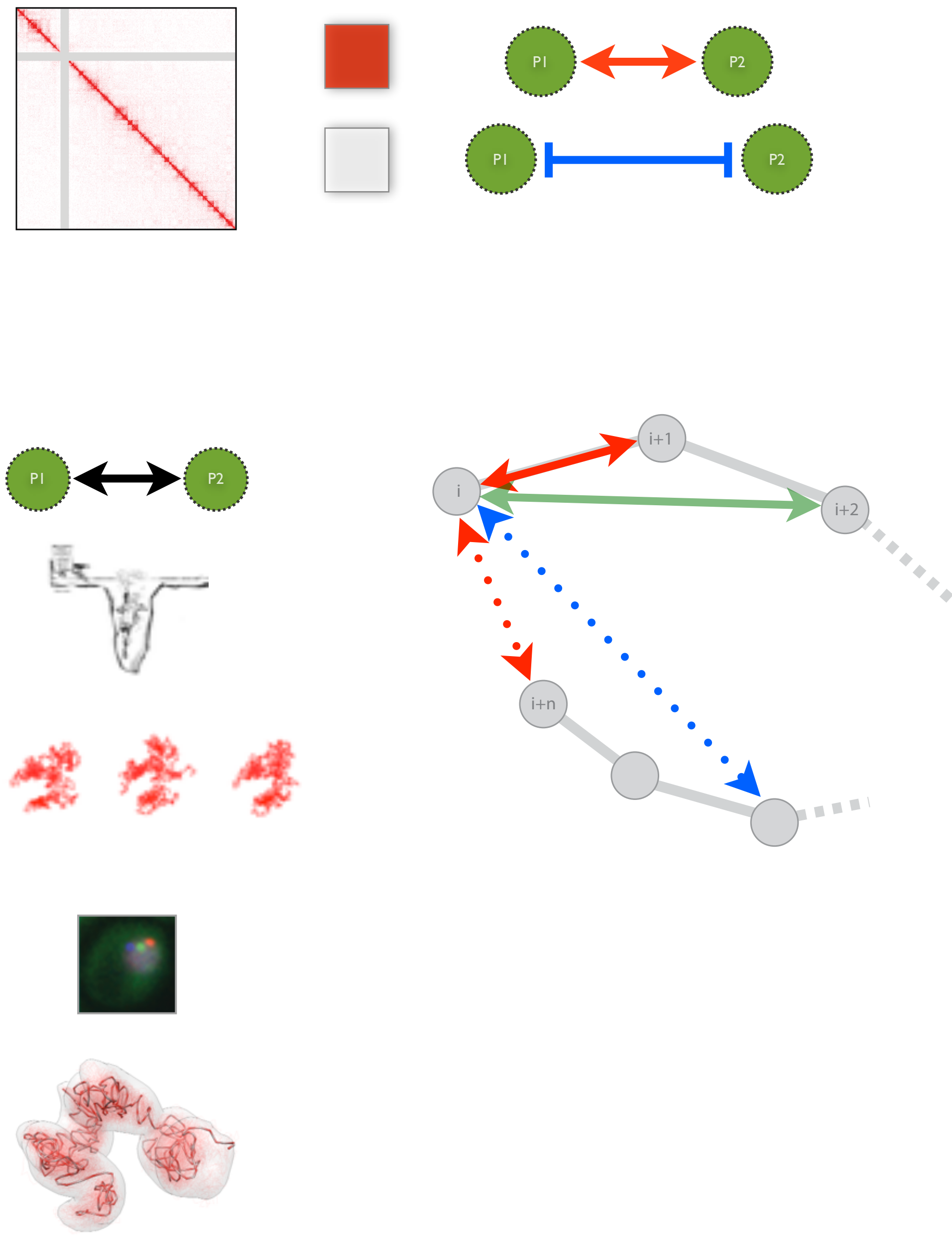
Experimental input and system representation
- 5C data
- Fragments and interaction maps

IMP Start

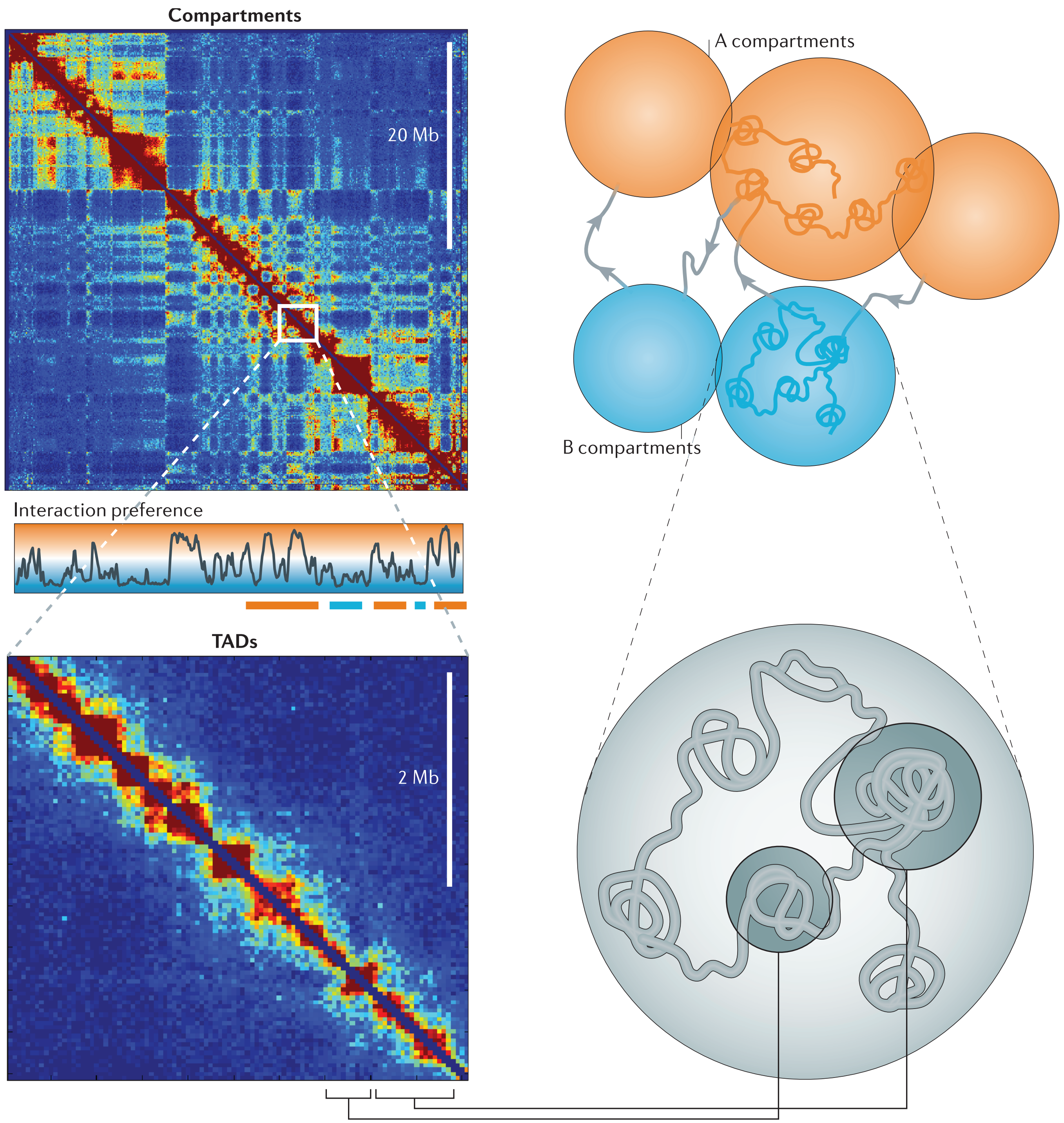
IMP sampler
- Restraints
- MonteCarlo & Simulating annealing
- Possible solitions

Results analysis
- Additional experimental data

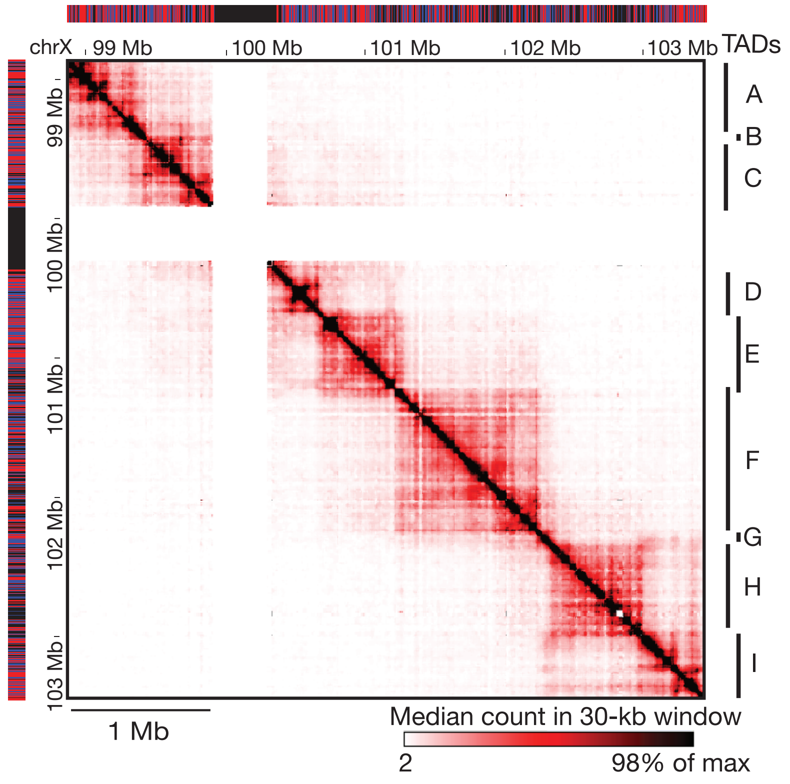
Final model
- Selection
- Genome organization

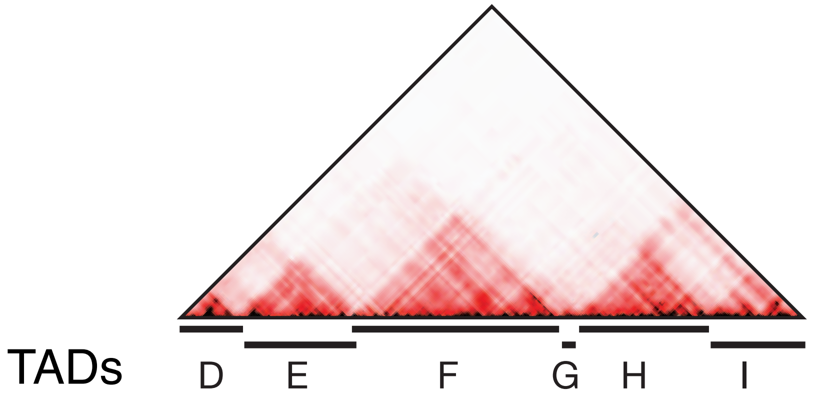


Genome compartments



Topologically Associating Domains (TADs)



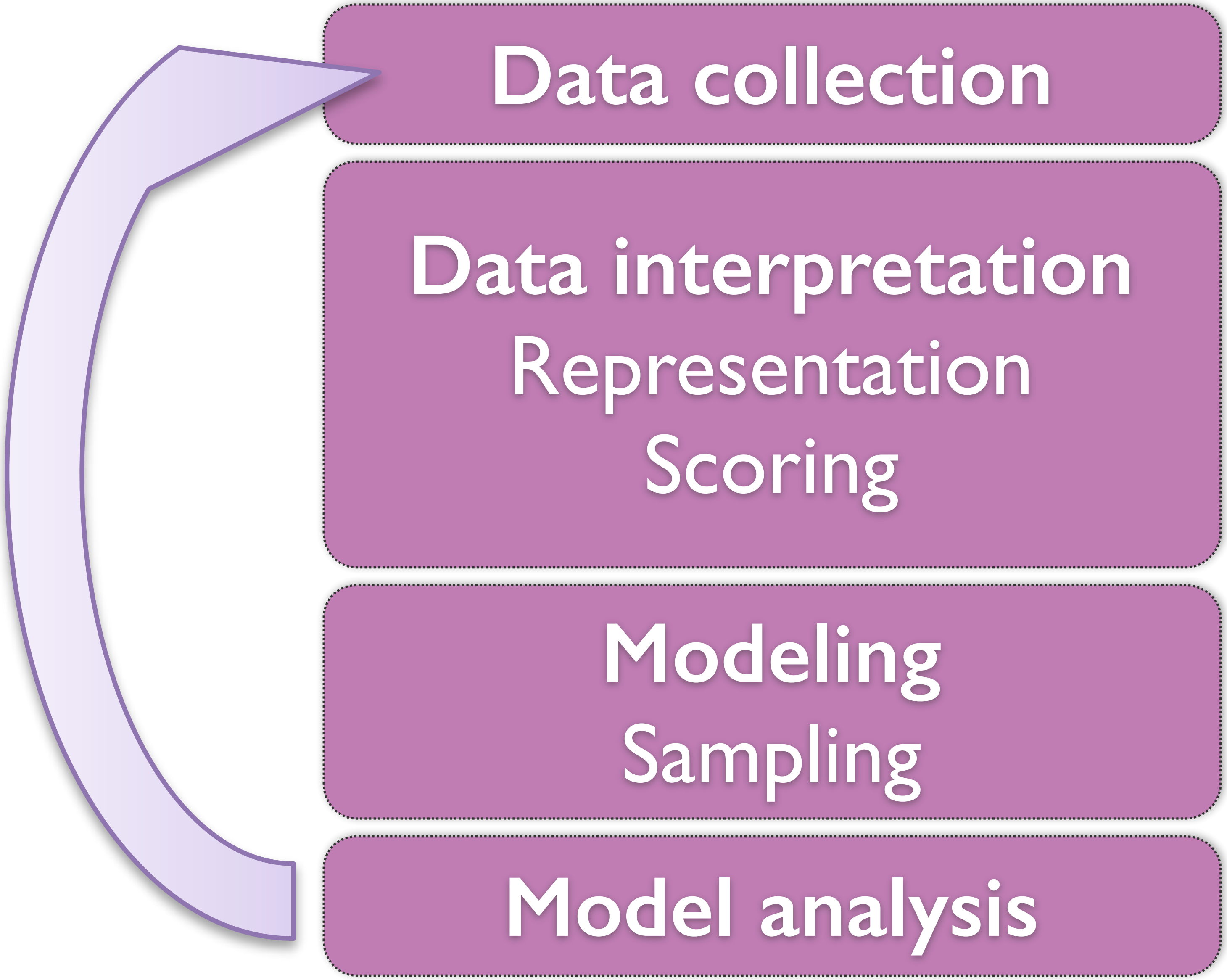


Topologically associating domains (TADs) can be made of up to hundreds of kb in size

Loci located within TADs tend to interact more frequently with each other than with loci located outside their domain

The human and mouse genomes are each composed of over 2,000 TADs, covering over 90% of the genome

Take home message



Data collection

Data interpretation
Representation
Scoring

Modeling
Sampling

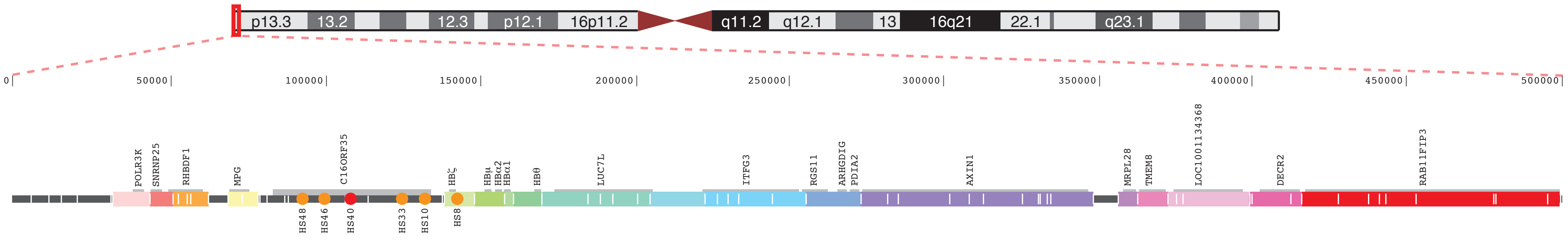
Model analysis

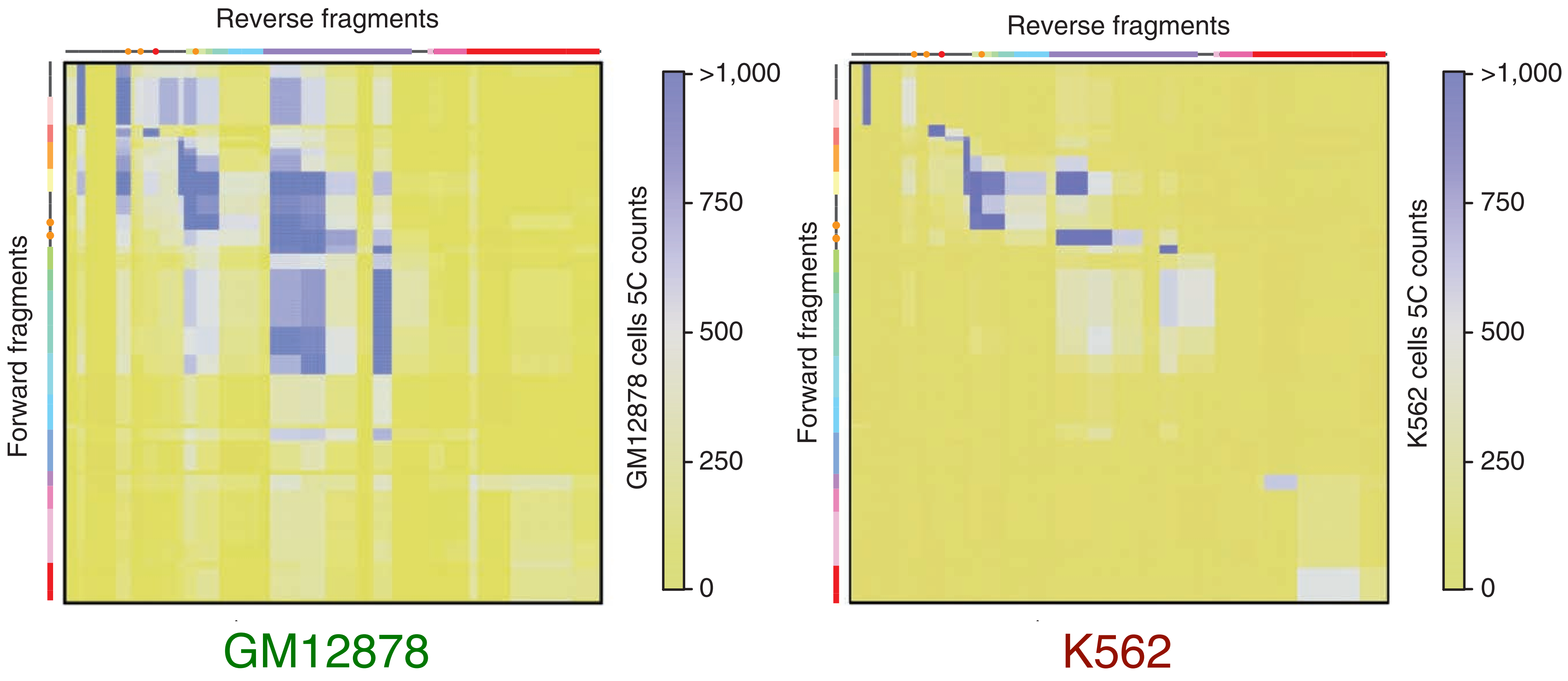
3C-like data: what can we get out of them?

Human α-globin domain

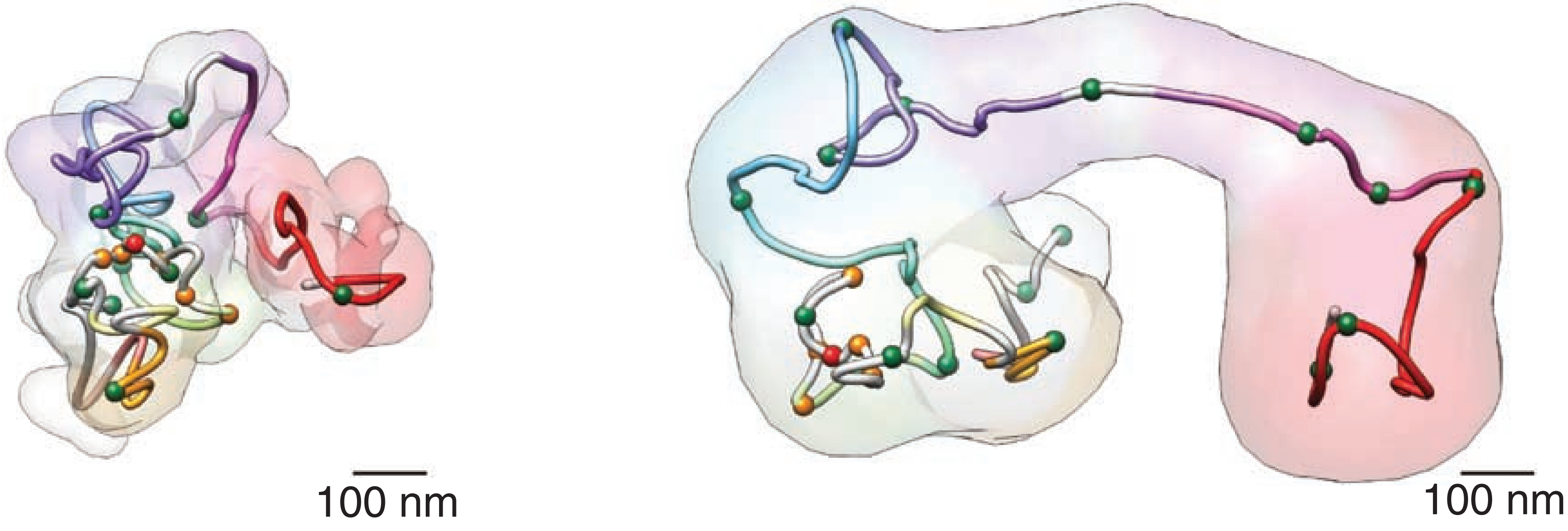
ENm008 genomic structure and environment

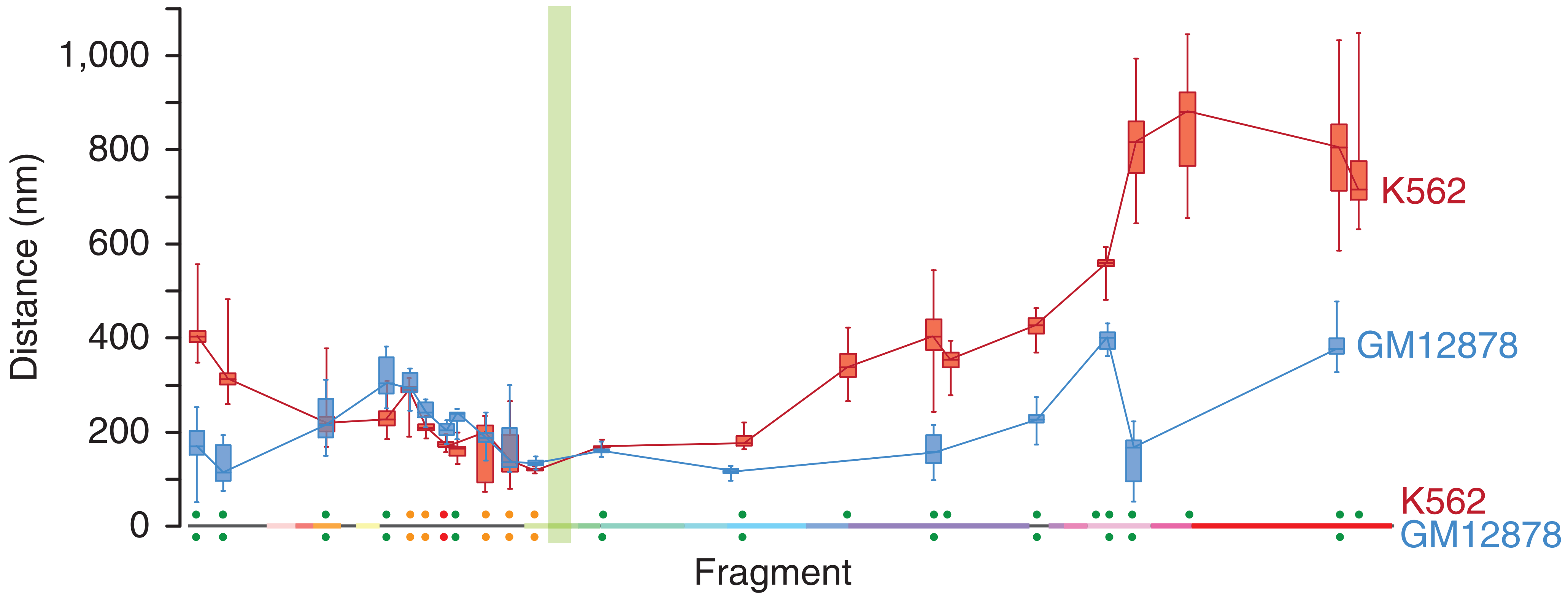
ENCODE Consortium. Nature (2007) vol. 447 (7146) pp. 799-816





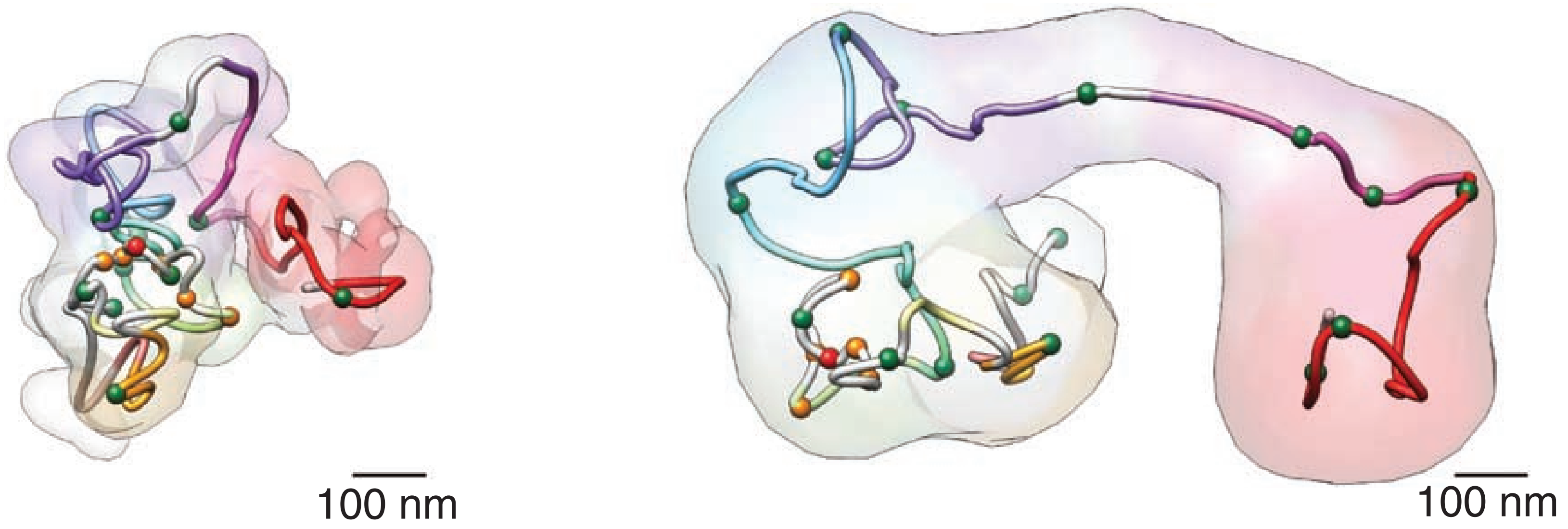
Long-range interactions

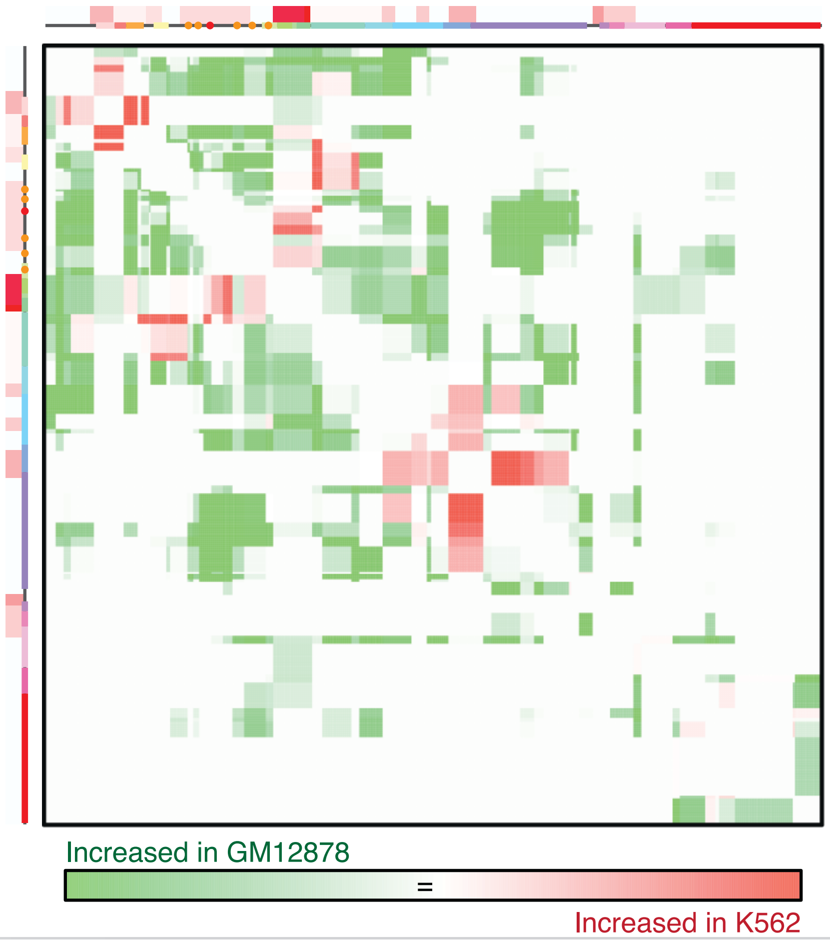




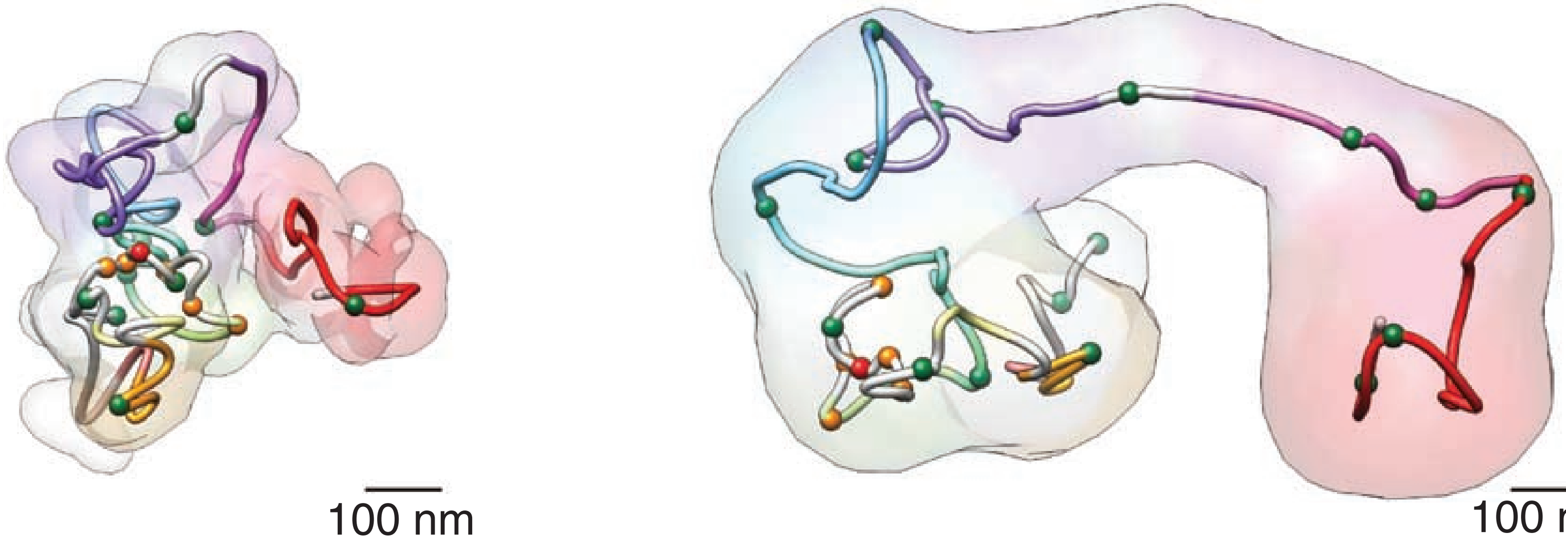
Chromatin globules

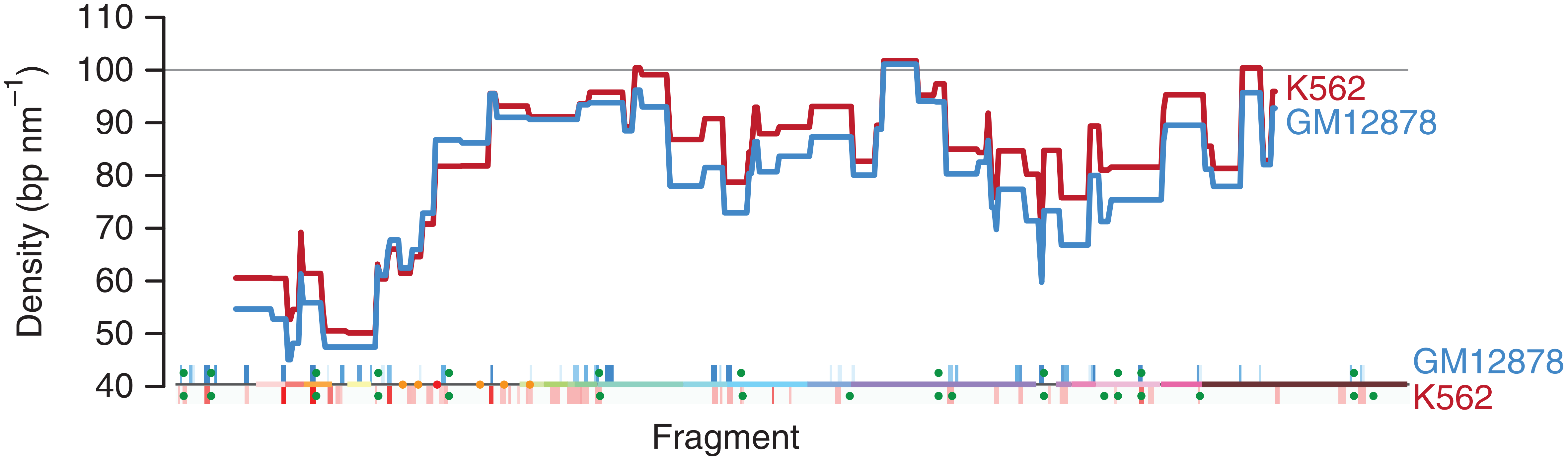
Frequency contact map differences



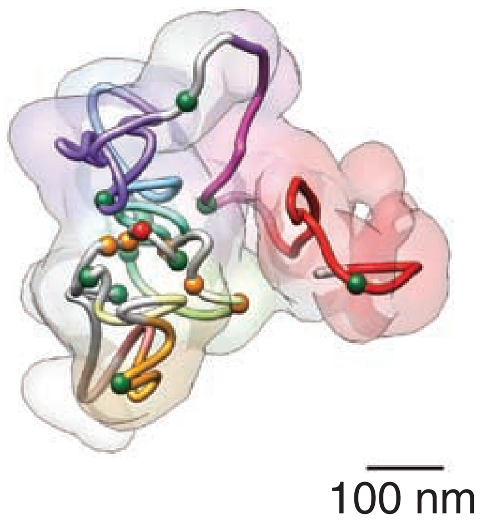


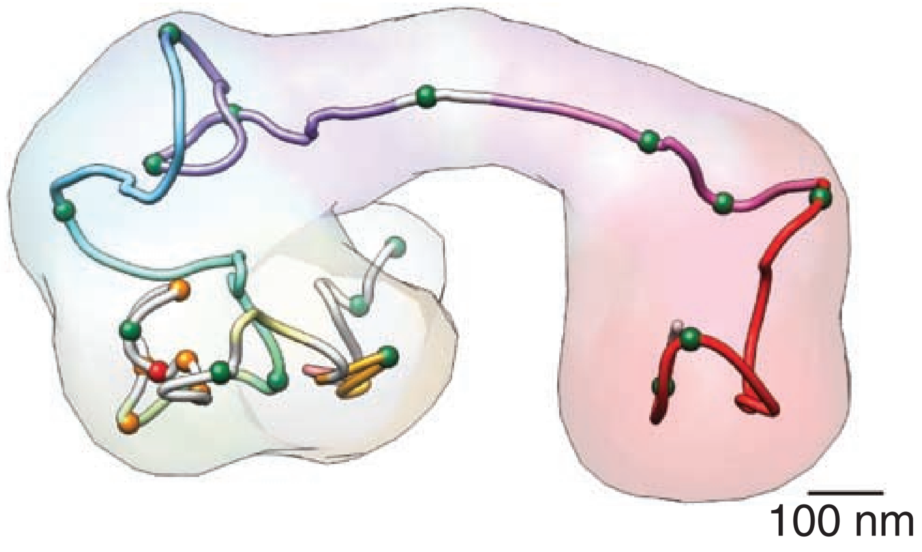
Chromatin compaction

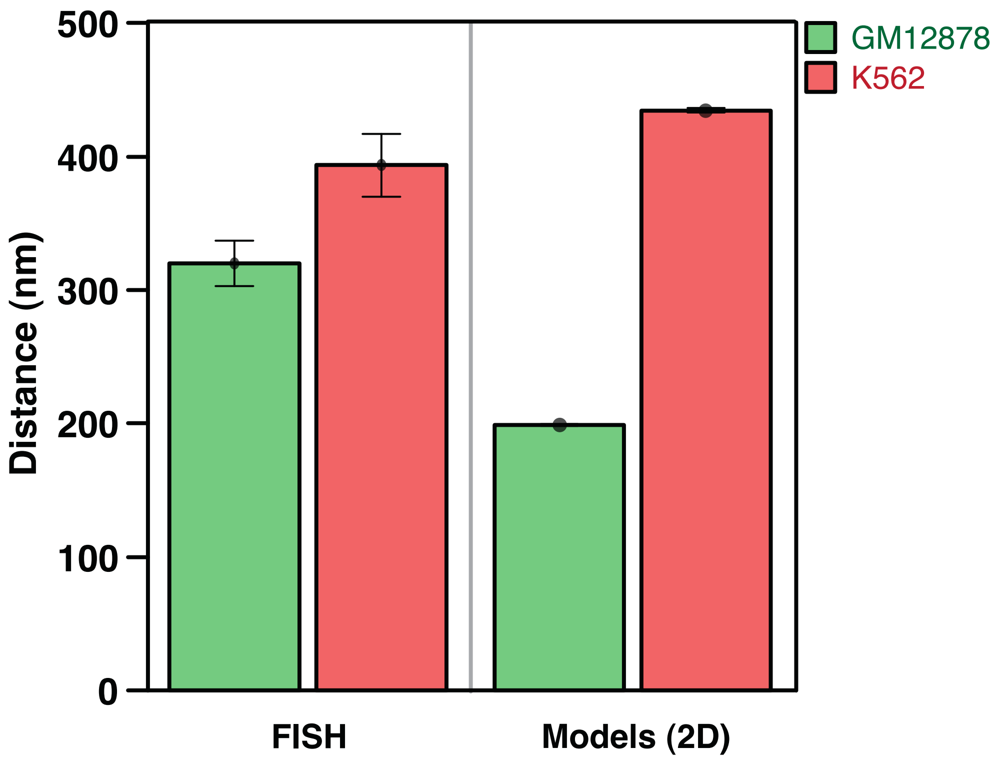


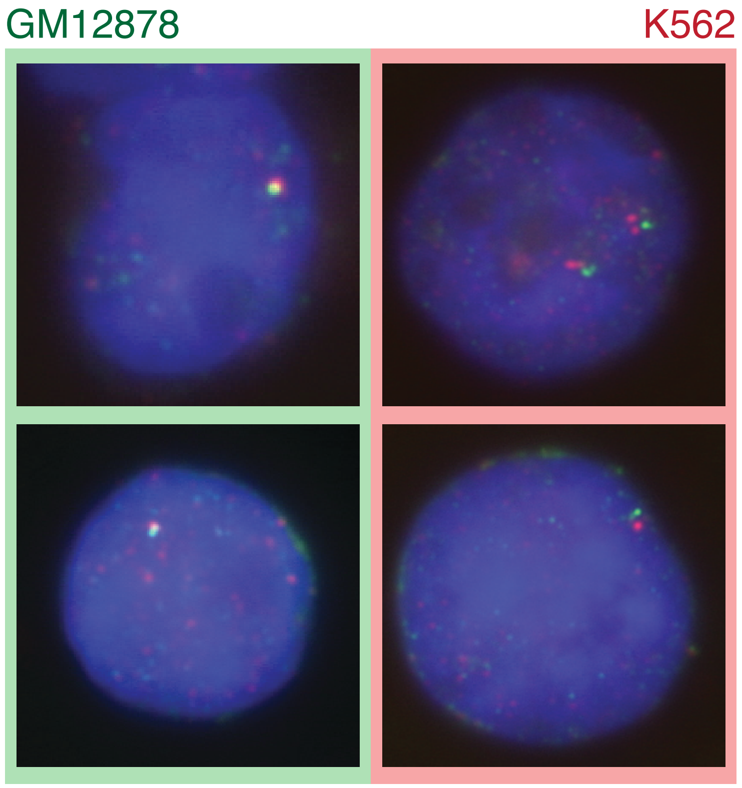


Model validation



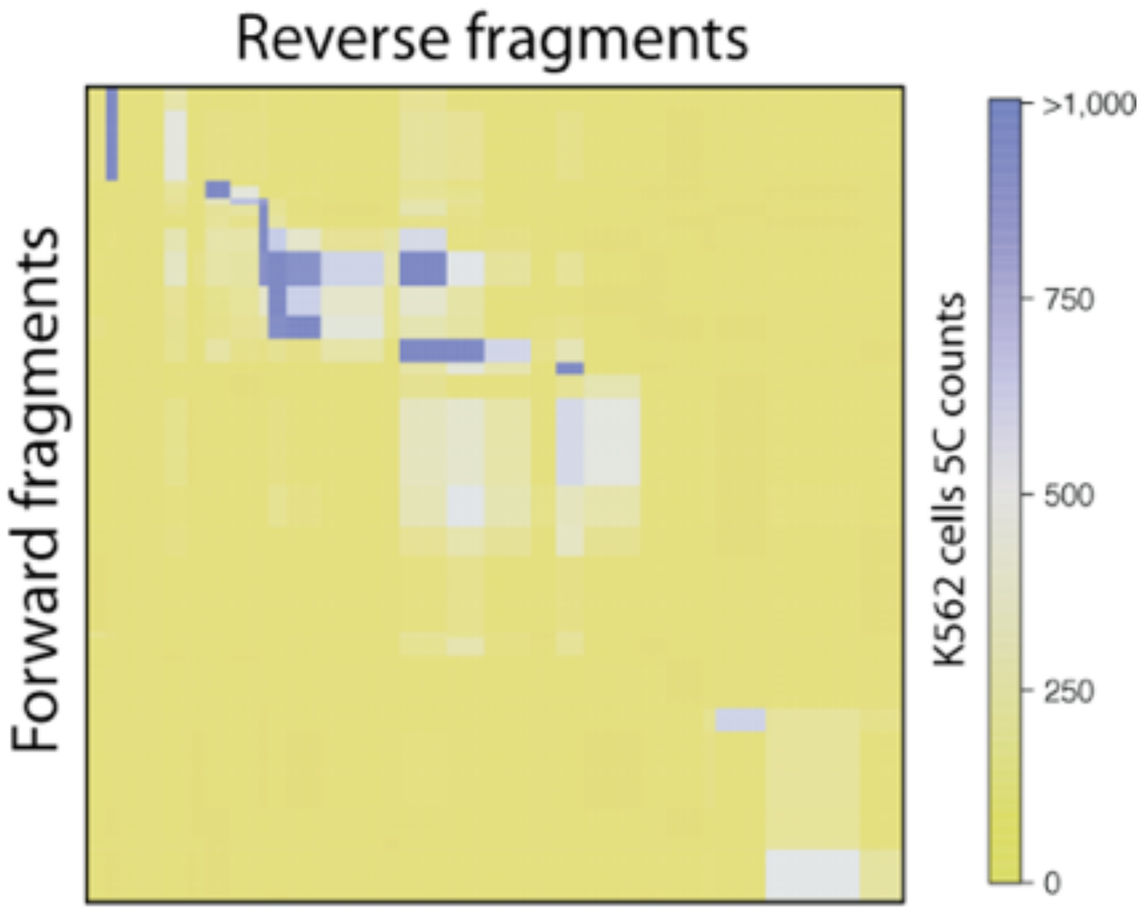


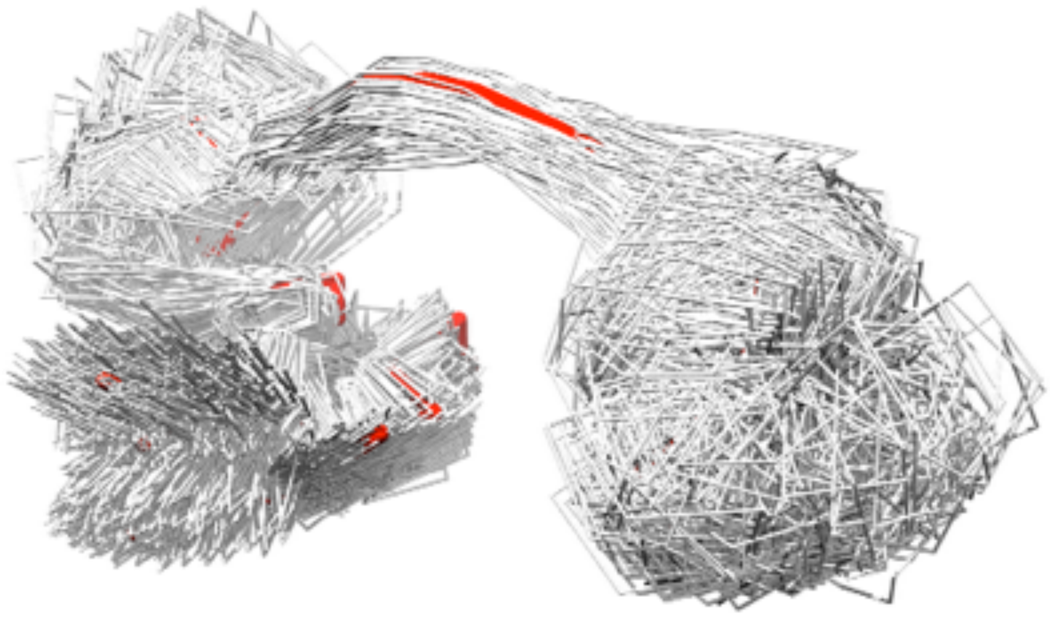




Summary

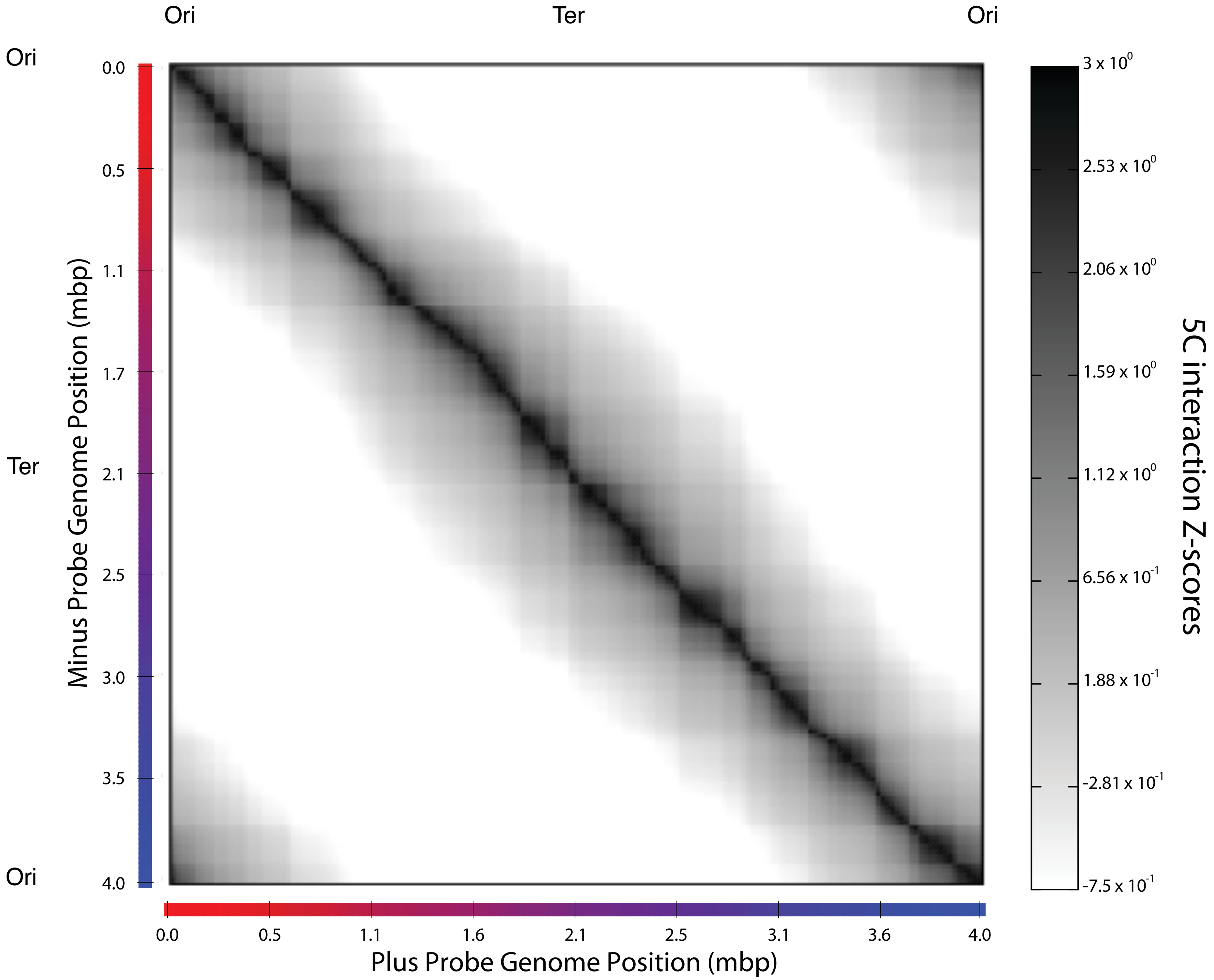
5C data results in consistent 3D models





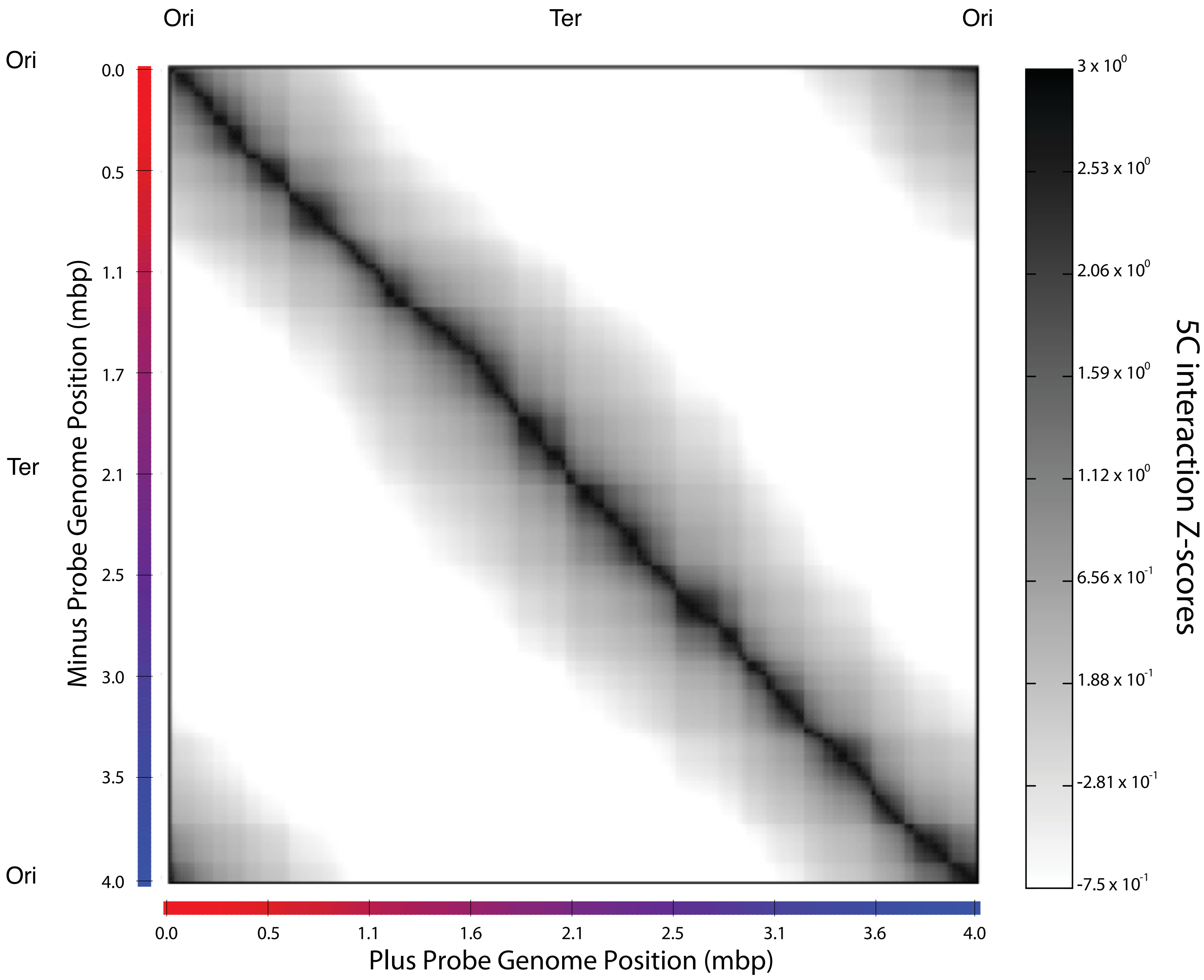
The genome of *Caulobacter crescentus*

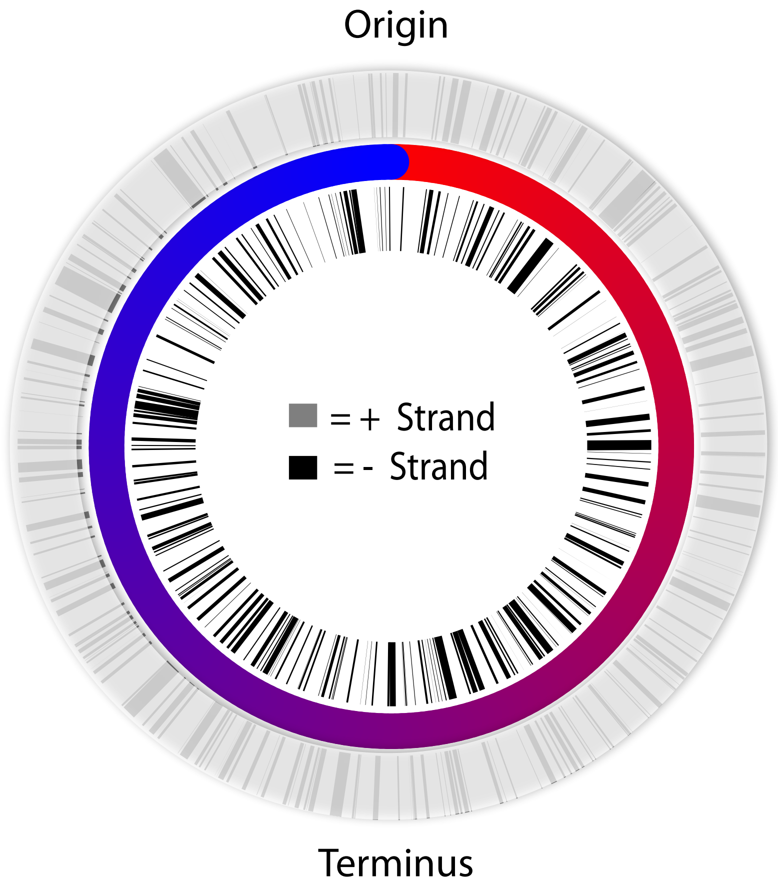
Toy interaction matrix



The genome of *Caulobacter crescentus*

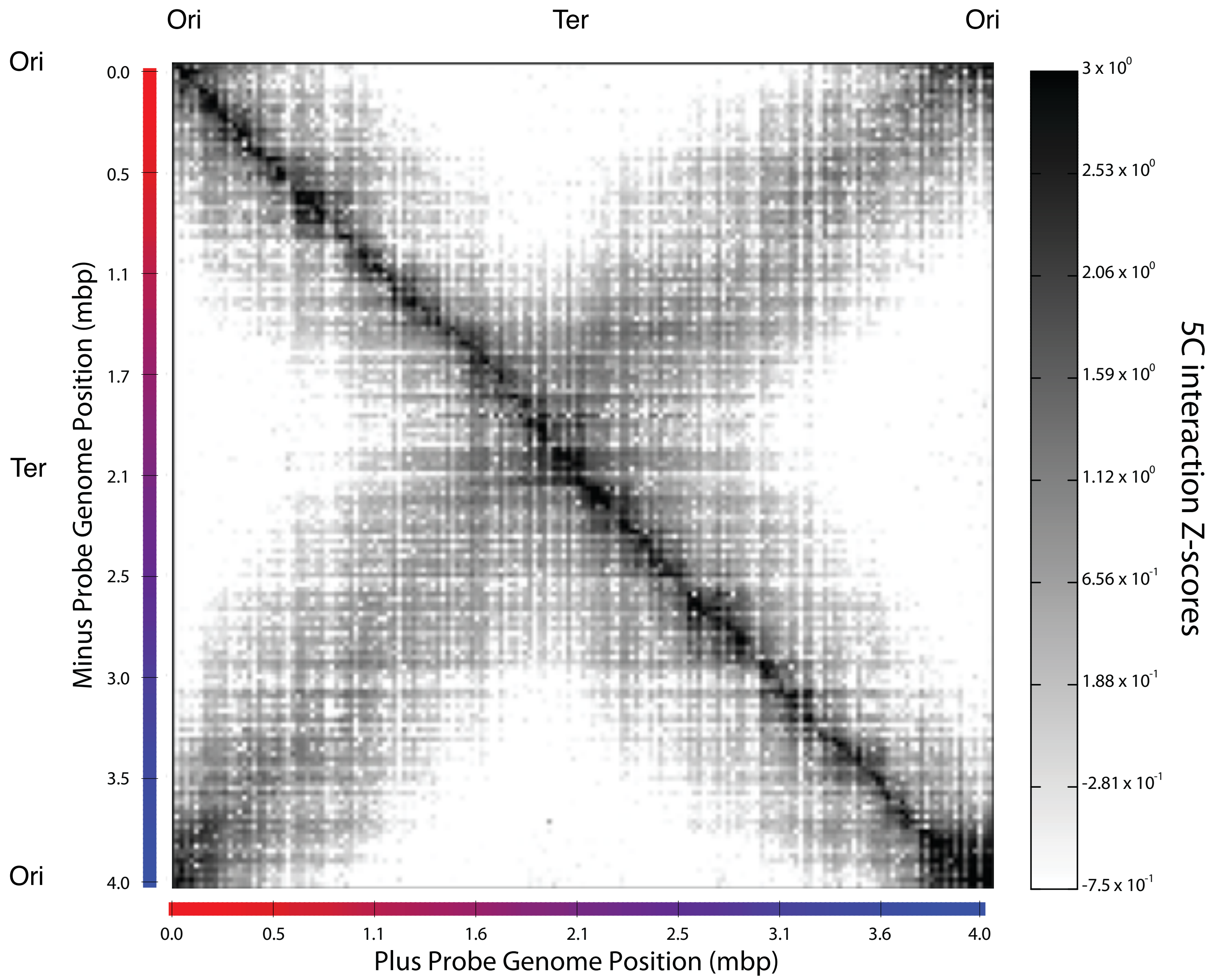
Toy interaction matrix





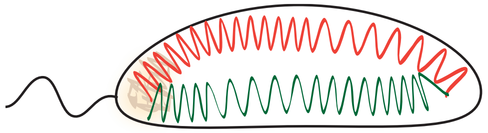
The genome of *Caulobacter crescentus*

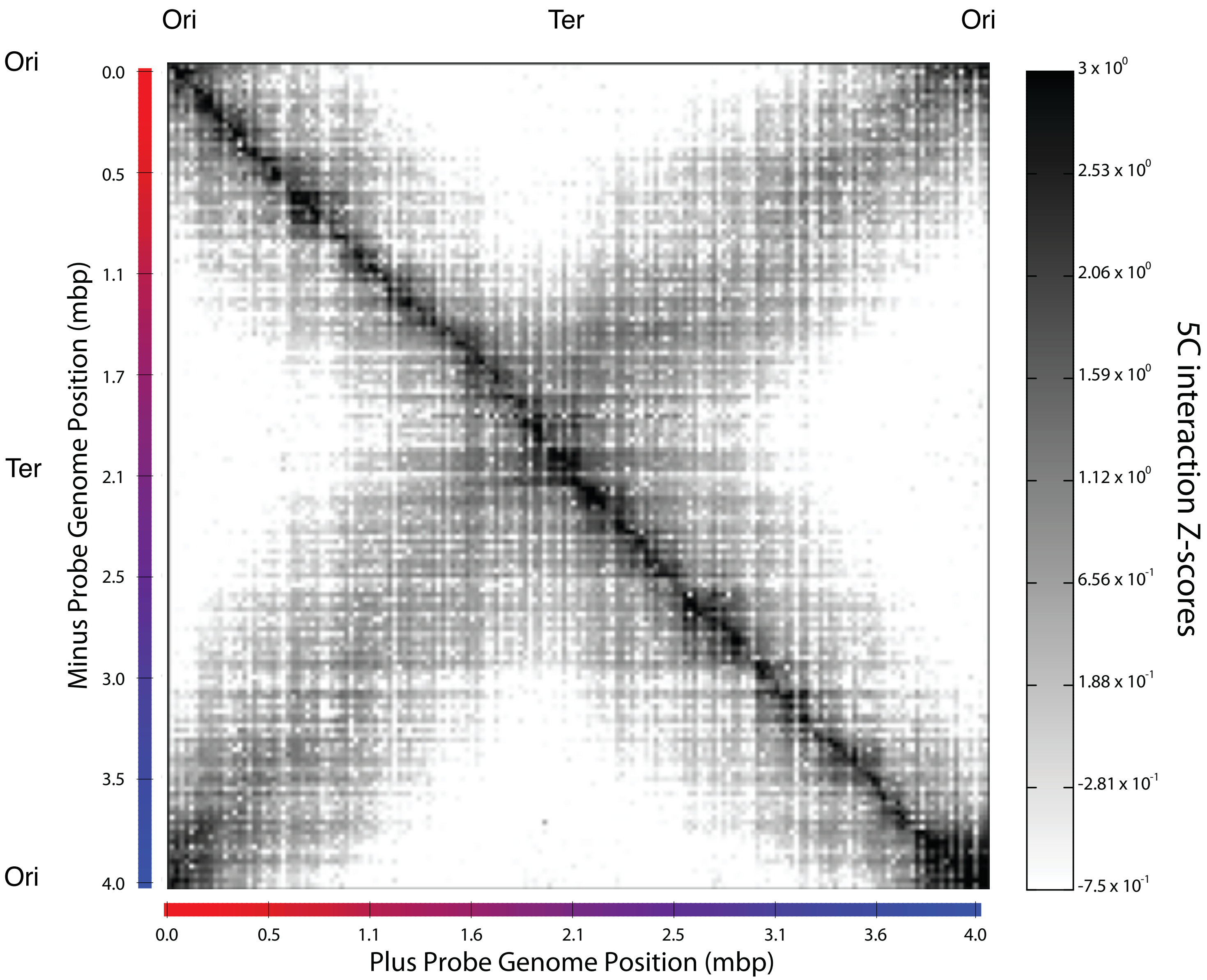
Real interaction matrix

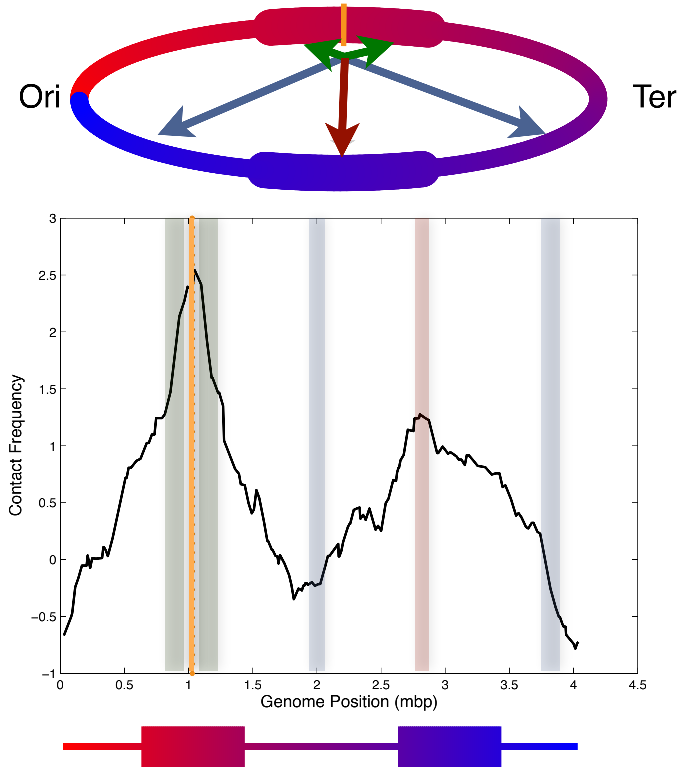


The genome of *Caulobacter crescentus*

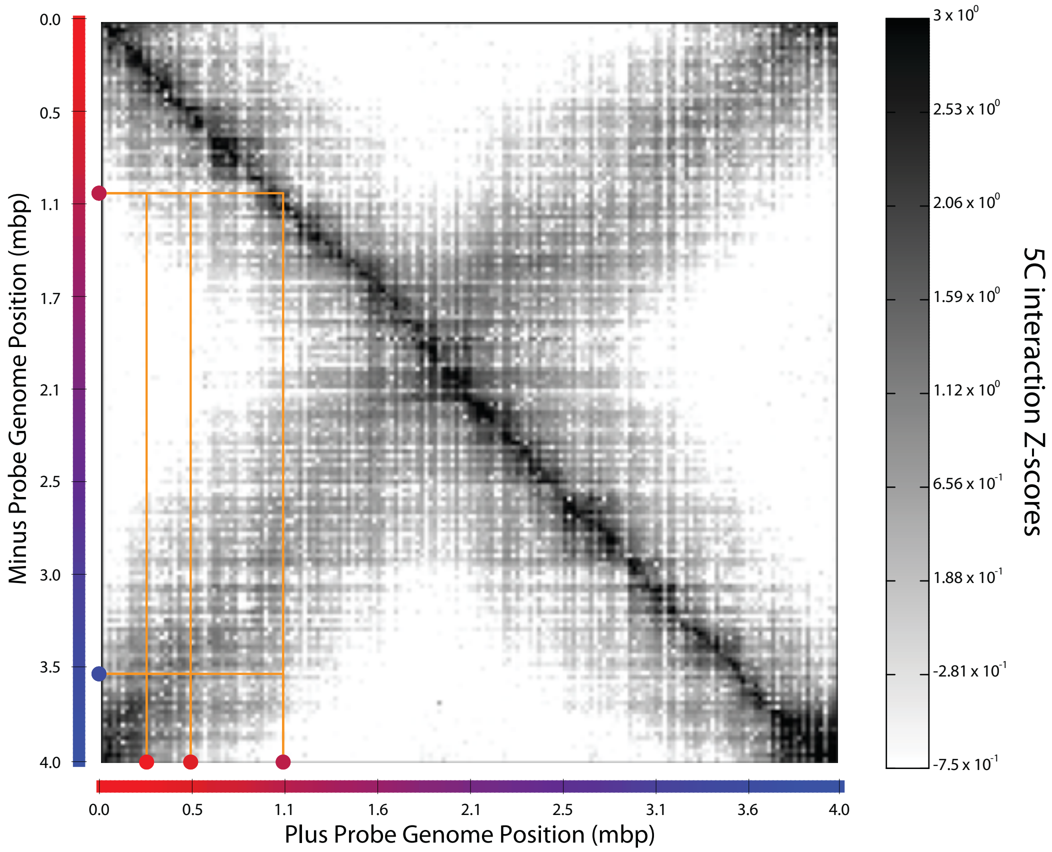
Real interaction matrix



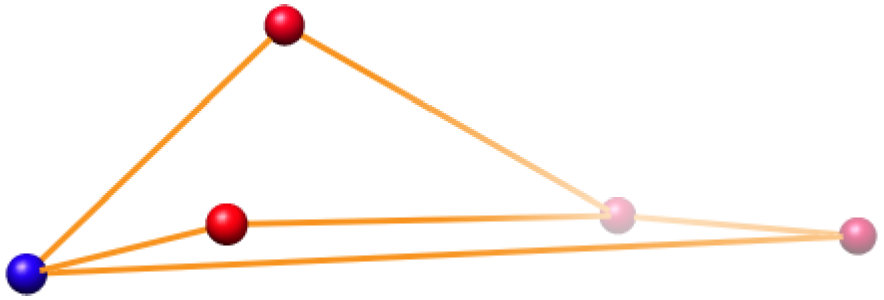


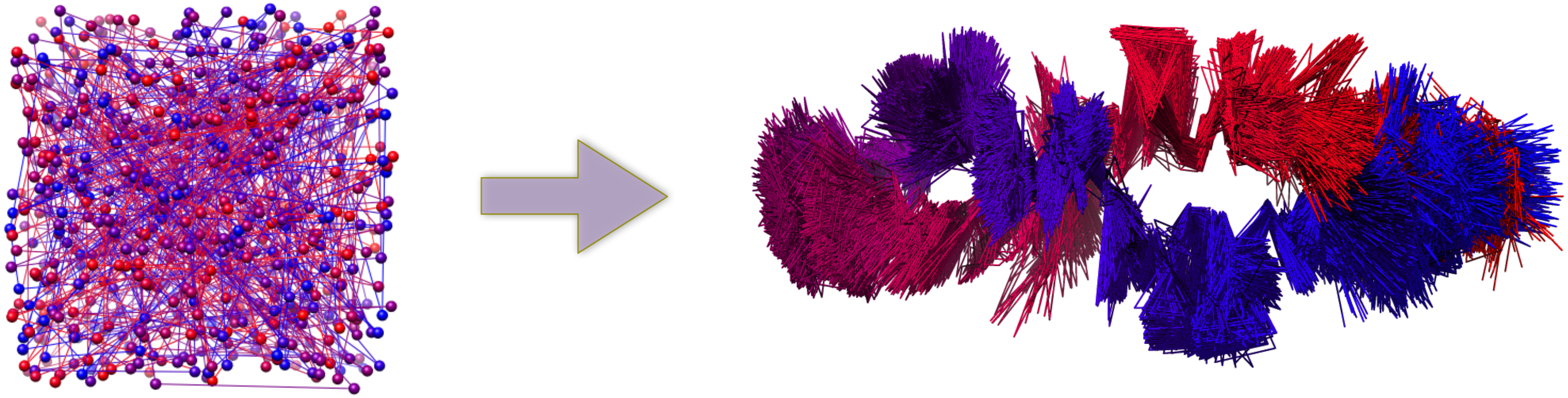


3D model building with the 5C + IMP approach



339 mers



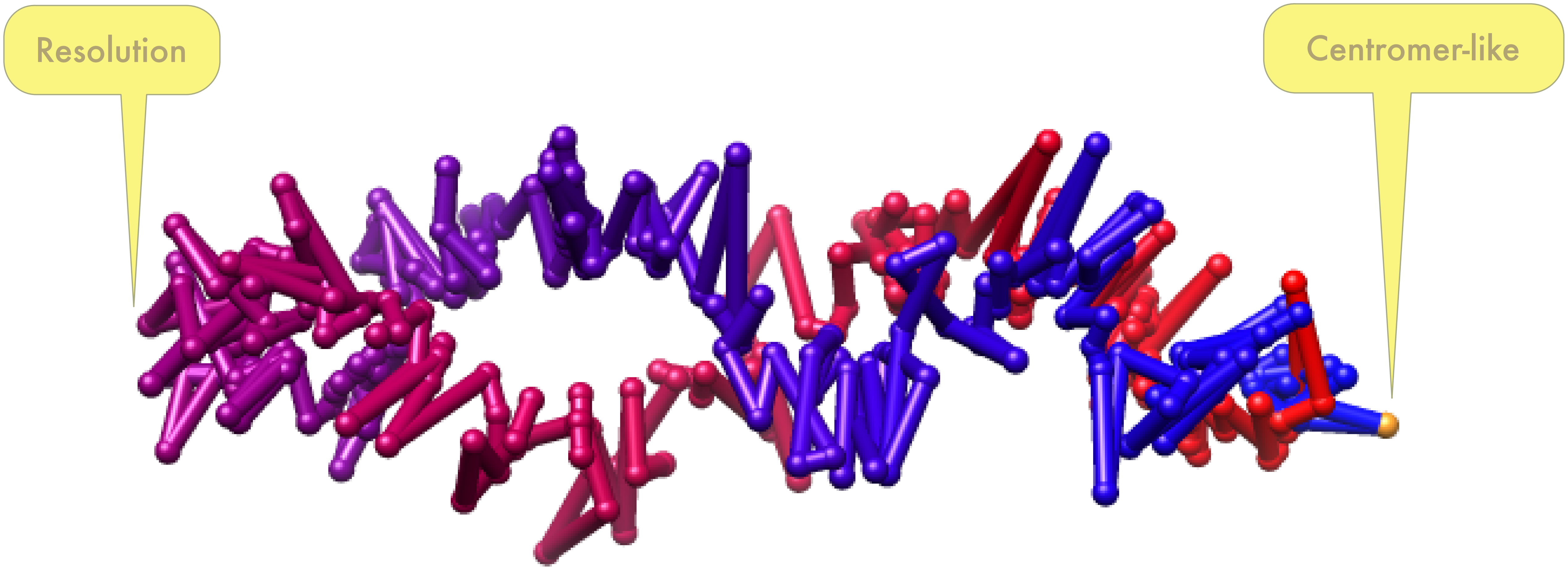


Genome organization in Caulobacter crescentus

Arms are helical

dif site 47±17Kb from Ter

parS sites 25±17Kb from Ori

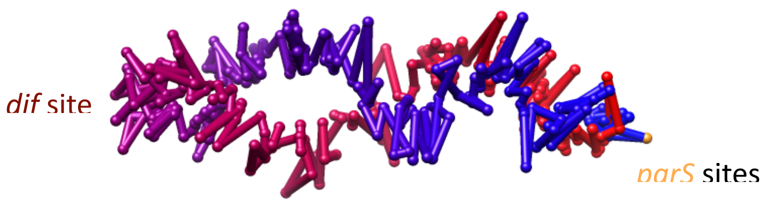


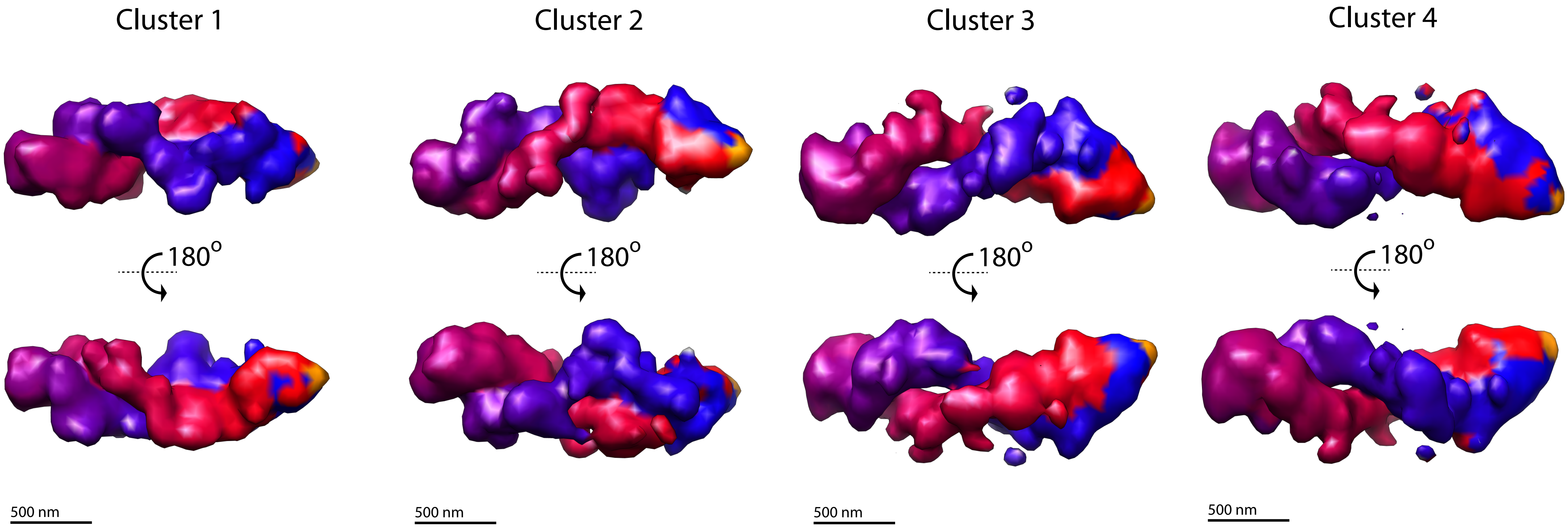
Resolution

Centromer-like

Genome organization in Caulobacter crescentus

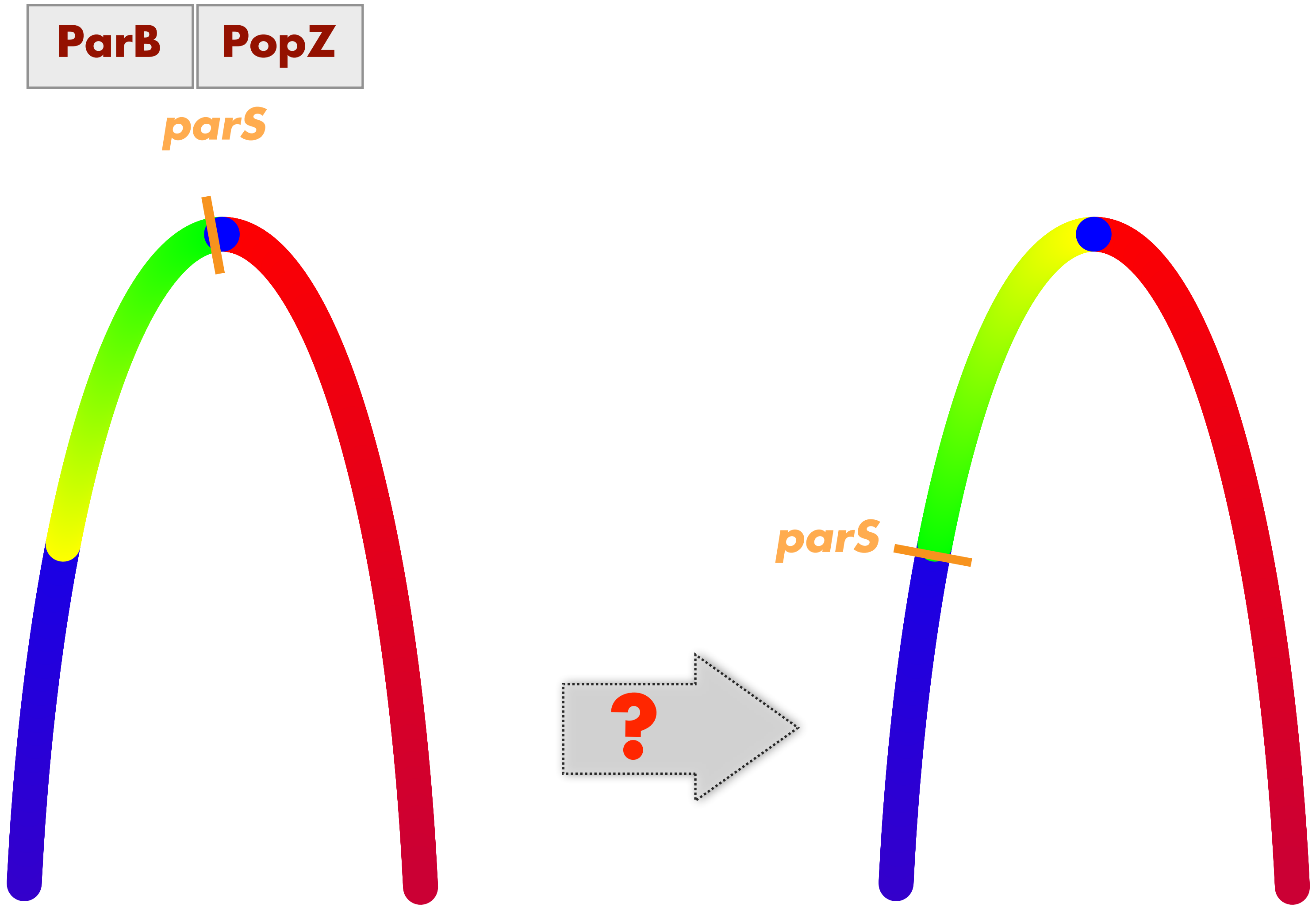
Arms are helical





MIRRORS!

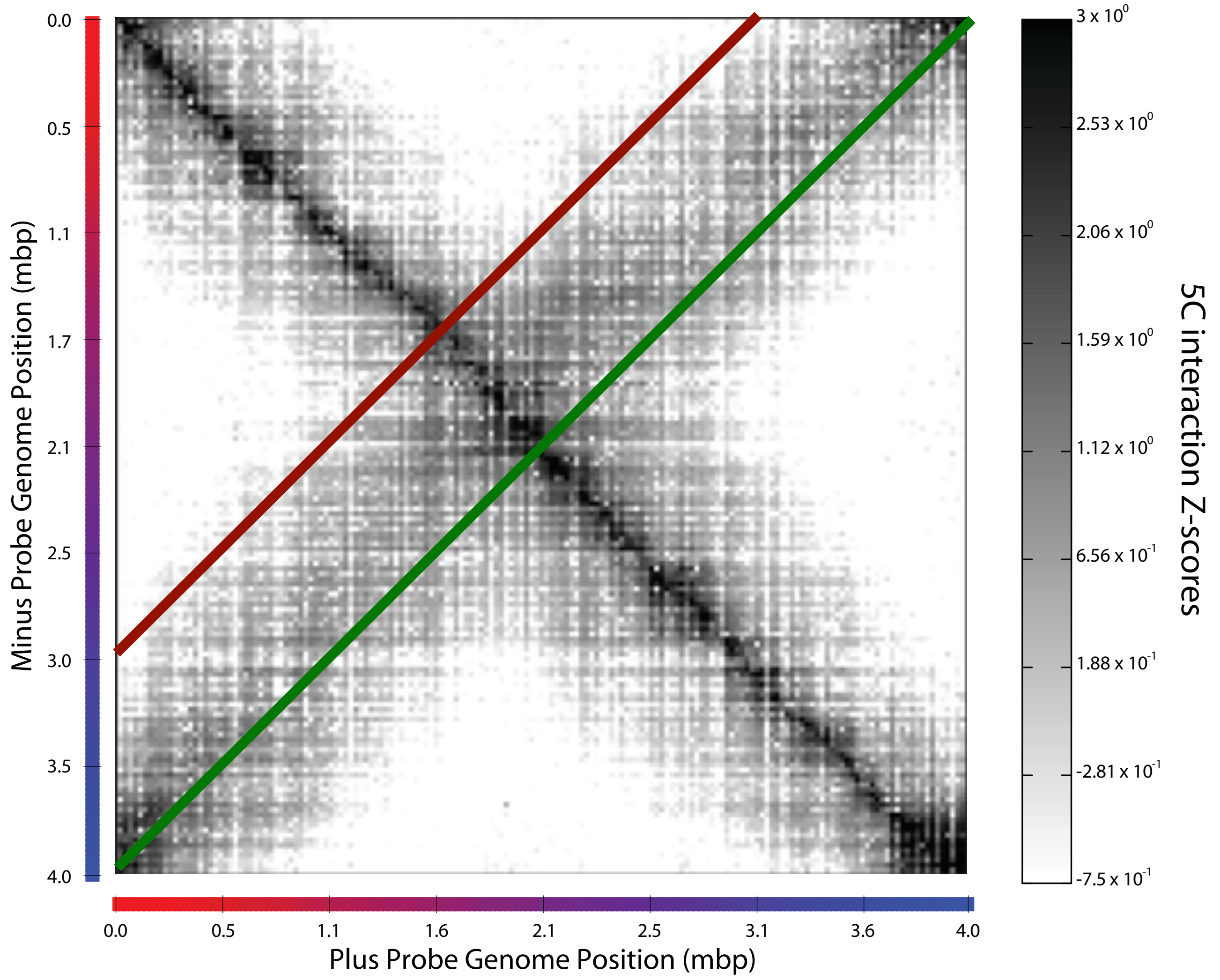
Moving the *parS* sites 400 Kb away from Ori

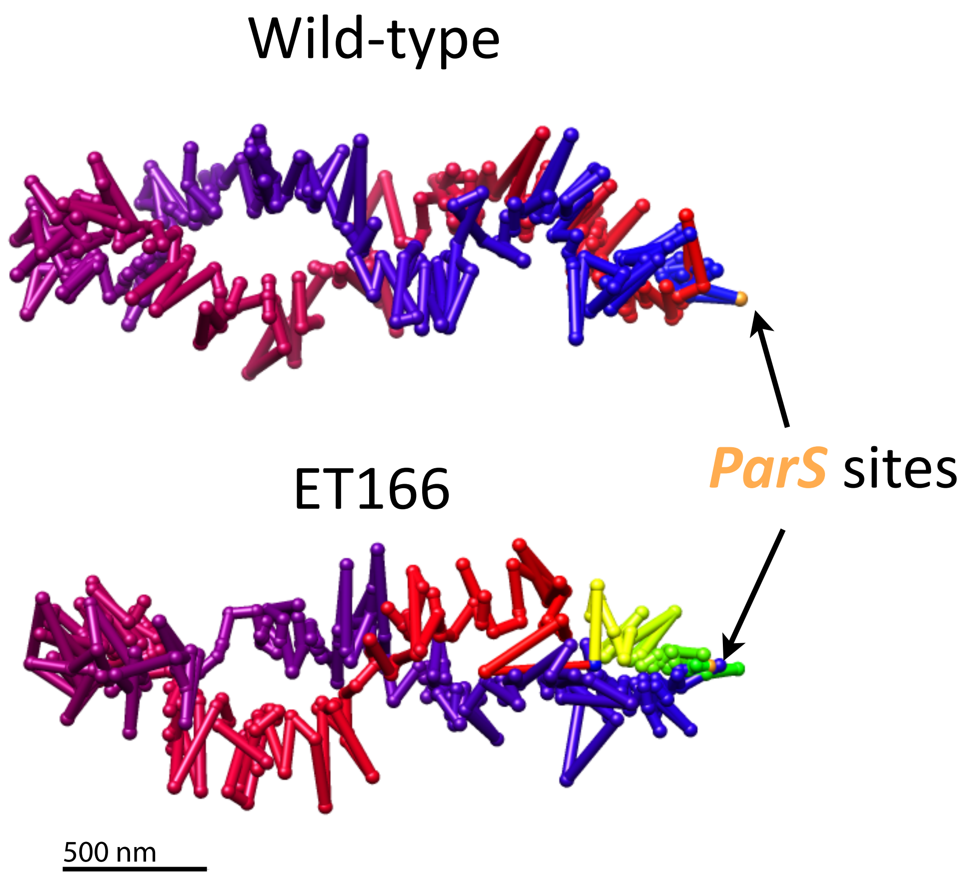


Wild-type

ET166

Moving the parS sites results in whole genome rotation!




Wild-type

ET166

Arms are **STILL** helical

Structure & function PRESERVED!!!

Take home message

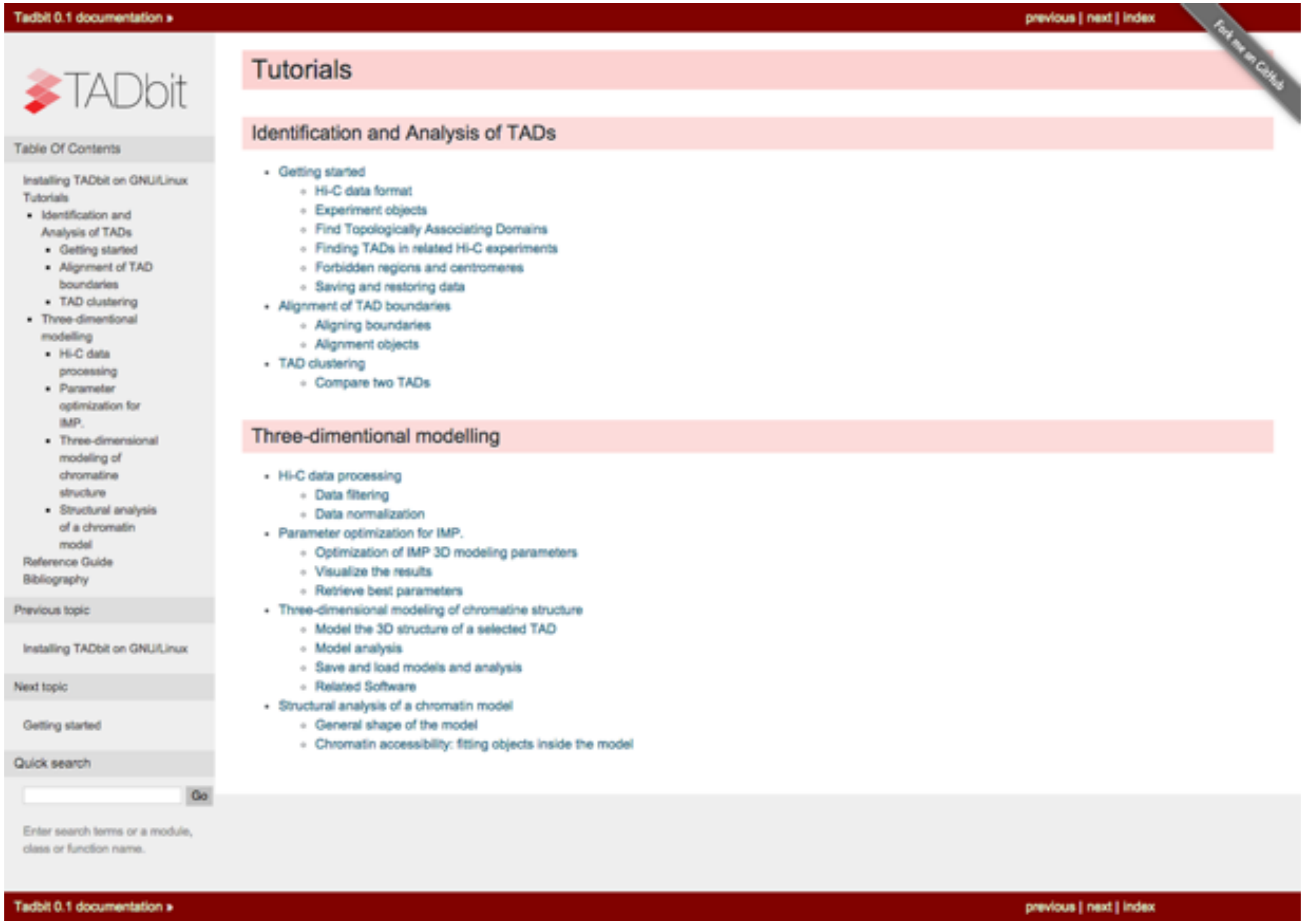
Chromatin = DNA + (histone) proteins

The genome is well organized and hierarchically packaged

Histone modifications affect chromatin structure and activity

3C-like data measure the frequency of interaction between distant loci

TADbit



Tutorials

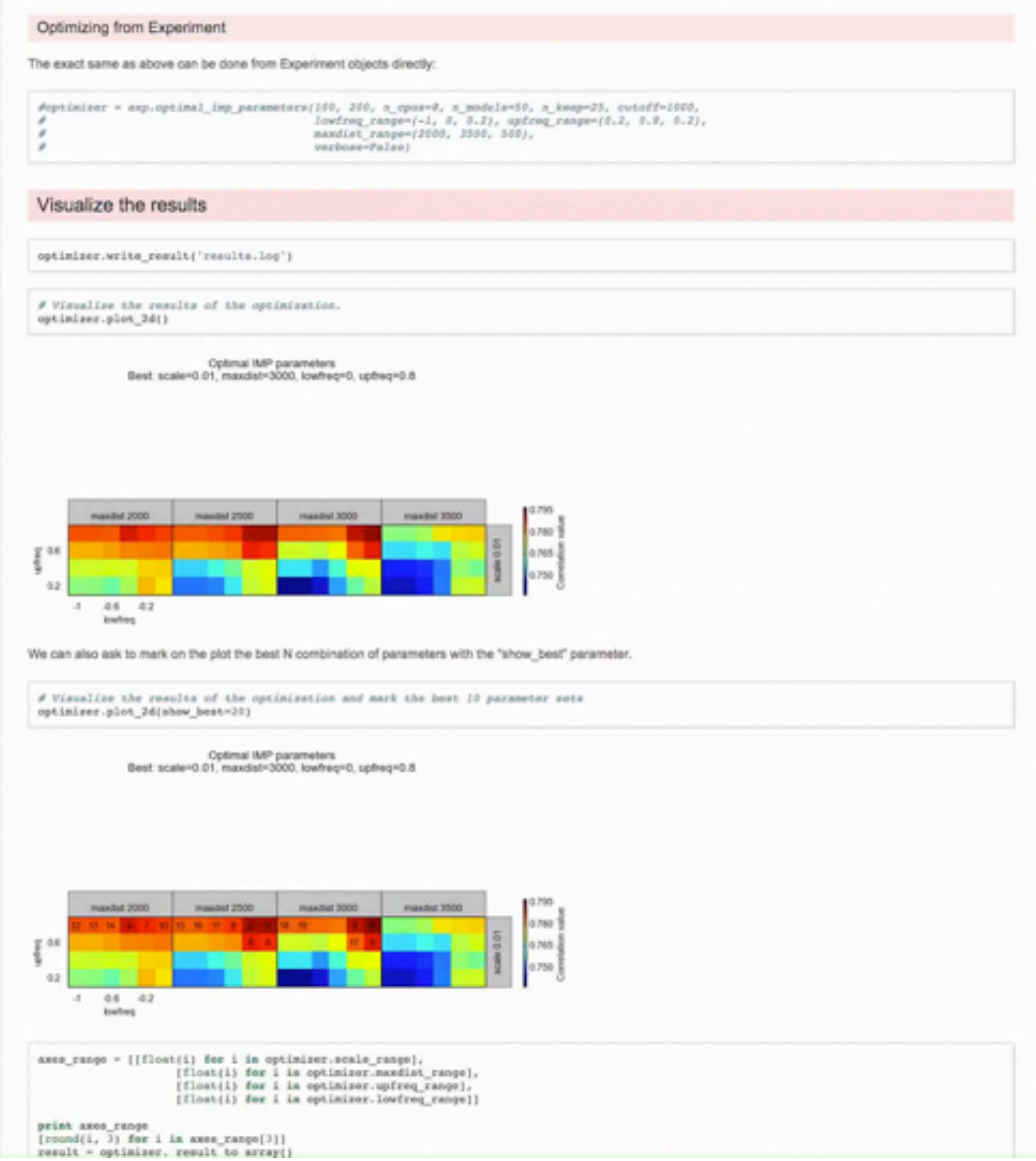
Identification and Analysis of TADs

- Getting started
  - Hi-C data format
  - Experiment objects
  - Find Topologically Associating Domains
  - Finding TADs in related Hi-C experiments
  - Forbidden regions and centromeres
  - Saving and restoring data
- Alignment of TAD boundaries
  - Aligning boundaries
  - Alignment objects
- TAD clustering
  - Compare two TADs

Three-dimentional modelling

- Hi-C data processing
  - Data filtering
  - Data normalization
- Parameter optimization for IMP.
  - Optimization of IMP 3D modeling parameters
  - Visualize the results
  - Retrieve best parameters
- Three-dimensional modeling of chromatine structure
  - Model the 3D structure of a selected TAD
  - Model analysis
  - Save and load models and analysis
  - Related Software
- Structural analysis of a chromatin model
  - General shape of the model
  - Chromatin accessibility: fitting objects inside the model

TADbit



Optimizing from Experiment

The exact same as above can be done from Experiment objects directly:

```
#optimizer = exp.optimal_imp_parameters(100, 200, n_cpus=8, n_models=50, n_keep=25, cutoff=1000,
#                                       lowfreq_range=(-1, 0, 0.2), upfreq_range=(0.2, 0.8, 0.2),
#                                       maxdist_range=(2000, 3500, 500),
#                                       verbose=False)
```

Visualize the results

```
optimizer.write_result('results.log')
```

```
# Visualize the results of the optimization.
optimizer.plot_2d()
```

Optimal IMP parameters
Best: scale=0.01, maxdist=3000, lowfreq=0, upfreq=0.8

We can also ask to mark on the plot the best N combination of parameters with the "show_best" parameter.

```
# Visualize the results of the optimization and mark the best 10 parameter sets
optimizer.plot_2d(show_best=20)
```

Optimal IMP parameters
Best: scale=0.01, maxdist=3000, lowfreq=0, upfreq=0.8

```
axes_range = [[float(i) for i in optimizer.scale_range],
              [float(i) for i in optimizer.maxdist_range],
              [float(i) for i in optimizer.upfreq_range],
              [float(i) for i in optimizer.lowfreq_range]]

print axes_range
[round(i, 3) for i in axes_range[3]]
result = optimizer. result to array()
```