Topologically associating domains of chromatin: methods and tools for calling

Part 1

Svyatoslav Sidorov

1The Dobzhansky Center for Genome Bioinformatics
St. Petersburg State University

Group meeting at BI
1. Introduction
Outline

1. Introduction

2. Topologically associating domains
1. Introduction

2. Topologically associating domains

3. TAD calling methods
1 Introduction

2 Topologically associating domains

3 TAD calling methods

4 Conclusion
Outline

1. Introduction

2. Topologically associating domains

3. TAD calling methods

4. Conclusion

5. Selected literature
1. Introduction

2. Topologically associating domains

3. TAD calling methods

4. Conclusion

5. Selected literature
Introduction

Introduction


**Question:** How is chromatin folded within euchromatin and heterochromatin compartments?

**Question:** How is chromatin folded within euchromatin and heterochromatin compartments?

**The answer** came with the development of chromatin conformation capture methods (3C, 2002; 4C, 2006; 5C, 2006; Hi-C, 2009).
Hi-C experiment scheme:

Lieberman-Aiden et al., 2009
Chromosome is split into $r$ bp bins ($r$ is called contact matrix resolution).

Contact matrix $C$ is built: $C(i, j) \equiv C(j, i)$ is a number of paired-end reads such that one read was mapped into bin $i$ and the other read was mapped into bin $j$. Contact matrix is usually represented as a heatmap.
Self-interacting domains can be seen on the main diagonal of a contact matrix (Dekker et al., 2013, adapted).
Dixon et al., 2012 found self-interacting domains in human and mouse using Hi-C data.
Dixon et al., 2012 found self-interacting domains in human and mouse using Hi-C data. They called such domains topologically associating domains (TADs). TAD is such a region that frequency of intra-TAD interactions is higher than inter-TAD interactions.
Topologically associating domains

- Dixon et al., 2012 found self-interacting domains in human and mouse using Hi-C data.

"Topological domains in mammalian genomes identified by analysis of chromatin interactions"

They called such domains **topologically associating domains (TADs)**. TAD is such a region that frequency of intra-TAD interactions is higher than inter-TAD interactions.

- Similar domains were found in *Drosophila* genome in the same year: Sexton et al., 2012; Hou et al., 2012.
Topologically associating domains

- Dixon et al., 2012 found self-interacting domains in human and mouse using Hi-C data.

**LETTER**

doi:10.1038/nature11082

Topological domains in mammalian genomes identified by analysis of chromatin interactions

Jesse R. Dixon¹²⁻³, Siddarth Selvaraj¹⁻⁴, Feng Yue¹, Audrey Kim¹, Yan Li¹, Yin Shen¹, Ming Hu⁵, Jun S. Liu⁵ & Bing Ren¹⁻⁶

- They called such domains topologically associating domains (TADs). TAD is such a region that frequency of intra-TAD interactions is higher than inter-TAD interactions.
- Similar domains were found in *Drosophila* genome in the same year: Sexton et al., 2012; Hou et al., 2012.
- TADs were also found in the same year in mouse X chromosome by Nora et al., 2012.
Topologically associating domains

- TADs are collections of many chromatin loops.
- TADs are separated by **TAD borders** (intervening chromatin).
- Mammalian TAD borders are enriched in active transcription, housekeeping genes, tRNA genes and SINE repeats, as well as binding sites for the architectural proteins CTCF and cohesin (**Dekker J. and Heard E., 2015**).
TAD-like domains were found in several organisms in 2012 – 2015 (Dekker J. and Heard E., 2015, adapted).
TADs as functional domains in mammals (Dekker J. and Heard E., 2015):
Topologically associating domains

**TADs as functional domains in mammals** *(Dekker J. and Heard E., 2015)*:

- TADs are units of coordinated gene expression.
TADs as functional domains in mammals (Dekker J. and Heard E., 2015):

- TADs are units of coordinated gene expression.
- Series of adjacent TADs correspond to replication domains.
Topologically associating domains

TADs as functional domains in mammals (Dekker J. and Heard E., 2015):

- TADs are units of coordinated gene expression.
- Series of adjacent TADs correspond to replication domains.
- Some TADs correspond to lamina-associated domains and other types of repressed chromatin.
TADs as functional domains in mammals (Dekker J. and Heard E., 2015):

- TADs are units of coordinated gene expression.
- Series of adjacent TADs correspond to replication domains.
- Some TADs correspond to lamina-associated domains and other types of repressed chromatin.
- Mammalian TAD borders are to a significant extent conserved between different cell types, and even between mouse and human.
Topologically associating domains

TADs as functional domains in mammals (Dekker J. and Heard E., 2015):

- TADs are units of coordinated gene expression.
- Series of adjacent TADs correspond to replication domains.
- Some TADs correspond to lamina-associated domains and other types of repressed chromatin.
- Mammalian TAD borders are to a significant extent conserved between different cell types, and even between mouse and human.
- Cell type-specific enhancers make loops with promoters of corresponding genes predominantly within TADs.
Topologically associating domains

TADs as functional domains in mammals (Dekker J. and Heard E., 2015):

- TADs are units of coordinated gene expression.
- Series of adjacent TADs correspond to replication domains.
- Some TADs correspond to lamina-associated domains and other types of repressed chromatin.
- Mammalian TAD borders are to a significant extent conserved between different cell types, and even between mouse and human.
- Cell type-specific enhancers make loops with promoters of corresponding genes predominantly within TADs.
- Internal interaction patterns of TADs are highly cell type-specific.
Topologically associating domains

**TADs as functional domains in mammals (Dekker J. and Heard E., 2015):**

- TADs are units of coordinated gene expression.
- Series of adjacent TADs correspond to replication domains.
- Some TADs correspond to lamina-associated domains and other types of repressed chromatin.
- Mammalian TAD borders are to a significant extent conserved between different cell types, and even between mouse and human.
- Cell type-specific enhancers make loops with promoters of corresponding genes predominantly within TADs.
- Internal interaction patterns of TADs are highly cell type-specific.
- TADs have hierarchical folding and consist of sub-TADs (Cubeñas-Potts C. and Corces V. G., 2015; Rao et al., 2014).
Topologically associating domains

TADs as functional domains in mammals (Dekker J. and Heard E., 2015):

- TADs are units of coordinated gene expression.
- Series of adjacent TADs correspond to replication domains.
- Some TADs correspond to lamina-associated domains and other types of repressed chromatin.
- Mammalian TAD borders are to a significant extent conserved between different cell types, and even between mouse and human.
- Cell type-specific enhancers make loops with promoters of correspondent genes predominantly within TADs.
- Internal interaction patterns of TADs are highly cell type-specific.
- TADs have hierarchical folding and consist of sub-TADs (Cubeñas-Potts C. and Corces V. G., 2015; Rao et al., 2014).

Self-interacting domains in other organisms can have different functions (Dekker J. and Heard E., 2015).
Outline

1. Introduction
2. Topologically associating domains
3. TAD calling methods
4. Conclusion
5. Selected literature
Let’s partition each chromosome into \( r \) bp bins, where \( r \) is a contact matrix resolution.
Let’s partition each chromosome into $r$ bp bins, where $r$ is a contact matrix resolution.

Contacts within the chromosome can then be visualized like this. Each arc denotes a pair of reads.
Let’s partition each chromosome into $r$ bp bins, where $r$ is a contact matrix resolution. Contacts within the chromosome can then be visualized like this. Each arc denotes a pair of reads. Then $A_L(i)$ is the number of read pairs that map from the bin $i$ to the upstream $L$ bp. $L$ should be a multiple of $r$. 
Let’s partition each chromosome into $r$ bp bins, where $r$ is a contact matrix resolution.

Contacts within the chromosome can then be visualized like this. Each arc denotes a pair of reads.

Then $A_L(i)$ is a number of read pairs that map from the bin $i$ to the upstream $L$ bp.

And $B_L(i)$ is a number of read pairs that map from the bin $i$ to the downstream $L$ bp.
At the end of a TAD we expect a bias in contact frequency towards upstream regions.
At the end of a TAD we expect a bias in contact frequency towards upstream regions.

And vice versa: at the beginning of a TAD we expect a bias in contact frequency towards downstream regions.
We can use this bias for TAD calling. Consider some bin $i$ and its $L$ bp vicinity. Let $A \equiv A_L(i)$, $B \equiv B_L(i)$, $D \equiv D_L(i)$, and $E \equiv E_L(i)$. Then, let’s define directionality index (Dixon et al., 2012)

$$DI = \frac{B - A}{|B - A|} \left( \frac{(A - E)^2}{E} + \frac{(B - E)^2}{E} \right),$$

where $E \equiv E_L(i) = \frac{A_L(i) + B_L(i)}{2}$ is an expected number of reads (without the upstream or downstream contact frequency bias).
We can use this bias for TAD calling. Consider some bin $i$ and its $L$ bp vicinity. Let $A \equiv A_L(i)$, $B \equiv B_L(i)$, $D \equiv D_L(i)$, and $E \equiv E_L(i)$. Then, let’s define directionality index (Dixon et al., 2012)

$$DI = \frac{B - A}{|B - A|} \left( \frac{(A - E)^2}{E} + \frac{(B - E)^2}{E} \right),$$

where $E \equiv E_L(i) = \frac{A_L(i) + B_L(i)}{2}$ is an expected number of reads (without the upstream or downstream contact frequency bias).

At the end of a TAD DI should have a local minimum, and immediately at the beginning of the next TAD DI should have a local maximum.
An illustration of this idea from Dixon et al., 2012 (Hi-C data for hESC – human embryonic stem cell line, some region of chr2):
DI calculation from a contact matrix (fig. is based on Crane et al., 2015):
DI calculation from a contact matrix (fig. is based on Crane et al., 2015):

\[ DI = \frac{\sum B - \sum A}{|\sum B - \sum A|} \left( \frac{(\sum A - E)^2}{E} + \frac{(\sum B - E)^2}{E} \right), \]

where \( E = \frac{\sum A + \sum B}{2} \), \( \sum A \) and \( \sum B \) are sums of elements in contact submatrices \( A \) and \( B \), respectively.
Now we can define a **Hidden Markov Model (HMM)** for TAD calling with DI (Dixon et al., 2012):

```
“Hidden” DB (State 1, 2, or 3) → “Hidden” DB (State 1, 2, or 3) → “Hidden” DB (State 1, 2, or 3)
Bin_{i-1} → Bin_{i} → Bin_{n}
```

- **“Hidden” DB (State 1, 2, or 3)**
- **Bin_{i-1}**
- **Bin_{i}**
- **Bin_{n}**

- **Mixture of Gaussians**
- **Bin_{i-1}**
- **Bin_{i}**
- **Bin_{n}**

- **“Observed” DI**
- **Bin_{i-1}**
- **Bin_{i}**
- **Bin_{n}**

**Labeling:**
- **“Upstream Bias” - State 1**
- **“Downstream Bias” - State 2**
- **No Bias - State 3**
Baum-Welch algorithm was used (somehow...) to compute maximum likelihood estimates of the model and the parameter estimates of transition and emission.
Baum-Welch algorithm was used (somehow...) to compute maximum likelihood estimates of the model and the parameter estimates of transition and emission.

Forward-backward algorithm was used to estimate posterior marginals, i.e., \( \Pr(Q_t = q \mid D_1 = d_1, D_2 = d_2, \ldots, D_n = d_n) \), where \( q \) is a hidden state, \( t \in \{1, \ldots, n\} \), \( d_1, d_2, \ldots, d_n \) are emission values.
Baum-Welch algorithm was used (somehow...) to compute maximum likelihood estimates of the model and the parameter estimates of transition and emission.

Forward-backward algorithm was used to estimate posterior marginals, i.e., $\Pr(Q_t = q | D_1 = d_1, D_2 = d_2, \ldots, D_n = d_n)$, where $q$ is a hidden state, $t \in \{1, \ldots, n\}$, $d_1, d_2, \ldots, d_n$ are emission values.

For each chromosome the authors tried to use 1 – 20 mixtures of Gaussians and chose one set with the best goodness of fit using the AIC criterion: $\text{AIC} = 2k - 2 \ln(L)$, where $k$ is the number of parameters in the model and $L$ is the maximum likelihood estimate.
**TAD calling:**

- TAD begins at the beginning of the first DB state in a series of DB states.
- **TAD calling:**
  - TAD begins at the beginning of the first DB state in a series of DB states.
  - TAD is continuous through all DB states in the series and then – through all the states in a UB series.
TAD calling:
- TAD begins at the beginning of the first DB state in a series of DB states.
- TAD is continuous through all DB states in the series and then – through all the states in a UB series.
- TAD ends in the last UB state in the series of UB states.
**Directionality index**

- **TAD calling:**
  - TAD begins at the beginning of the first DB state in a series of DB states.
  - TAD is continuous through all DB states in the series and then – through all the states in a UB series.
  - TAD ends at the end of the last UB state in the series of UB states.

- **TAD borders:** a region between TADs is called **topological boundary** if its length is less than 400 kbp, otherwise it is called **unrecognized chromatin**.
• **TAD calling:**
  - TAD begins at the beginning of the first DB state in a series of DB states.
  - TAD is continuous through all DB states in the series and then – through all the states in a UB series.
  - TAD ends at the end of the last UB state in the series of UB states.

• **TAD borders:** a region between TADs is called **topological boundary** if its length is less than 400 kbp, otherwise it is called **unrecognized chromatin**.

• Topological boundaries in mouse ESC were found to be quite small, 76.33% of them being less than 50 kbp.
The main biological results in Dixon et al., 2012 are as follows:

- TADs were called in mouse and human ESC, as well as in some terminally differentiated cell types. E.g., about 91% of the mouse ESC is occupied by TADs with median size around 880 kbp.
The main biological results in Dixon et al., 2012 are as follows:

- TADs were called in mouse and human ESC, as well as in some terminally differentiated cell types. E.g., about 91% of the mouse ESC is occupied by TADs with median size around 880 kbp.
- TADs are stable across different cell types and highly conserved across species.
The main biological results in Dixon et al., 2012 are as follows:

- TADs were called in mouse and human ESC, as well as in some terminally differentiated cell types. E.g., about 91% of the mouse ESC is occupied by TADs with median size around 880 kbp.
- TADs are stable across different cell types and highly conserved across species.
- TAD borders are enriched for CTCF, housekeeping genes, tRNAs, and SINE retrotransposons.
The main biological results in Dixon et al., 2012 are as follows:

- TADs were called in mouse and human ESC, as well as in some terminally differentiated cell types. E.g., about 91% of the mouse ESC is occupied by TADs with median size around 880 kbp.
- TADs are stable across different cell types and highly conserved across species.
- TAD borders are enriched for CTCF, housekeeping genes, tRNAs, and SINE retrotransposons.

These results (and raw Hi-C data from the paper) are used in biological studies (see, e.g., Battulin et al., 2015, Rao et al., 2014, Van Bortle, 2014, Pope et al, 2014, Duggal et al., 2014, Kolovos et al., 2014, Zhao et al., 2013, Lu et al, 2013)
The main biological results in Dixon et al., 2012 are as follows:

- TADs were called in mouse and human ESC, as well as in some terminally differentiated cell types. E.g., about 91% of the mouse ESC is occupied by TADs with median size around 880 kbp.
- TADs are stable across different cell types and highly conserved across species.
- TAD borders are enriched for CTCF, housekeeping genes, tRNAs, and SINE retrotransposons.

Although Dixon et al., 2012 didn’t publish their scripts (they used MATLAB) and detailed description of the HMM, directionality index (DI) became a popular metric for TAD calling.
Although Dixon et al., 2012 didn’t publish their MATLAB scripts and detailed description of the HMM, directionality index (DI) became a popular metric for TAD calling. E. g.:

- **Pope et al., 2014** called TAD borders (without HMM) in human fibroblasts IMR90 in order to compare them to those previously called in Dixon et al., 2012 (higher resolution Hi-C data were used) and to use them in replication-timing studies.
Although Dixon et al., 2012 didn’t publish their MATLAB scripts and detailed description of the HMM, directionality index (DI) became a popular metric for TAD calling. E. g.:

- **Pope et al., 2014** called TAD borders (without HMM) in human fibroblasts IMR90 in order to compare them to those previously called in Dixon et al., 2012 (higher resolution Hi-C data were used) and to use them in replication-timing studies.

- **Dileep et al., 2015** calculated DI in six regions at several time points in the G1-phase of mouse mammary epithelial cell line (C127) watching a switch from a negligible to strong directionality bias that suggested formation of TADs.
Insulation score
Insulation score
**Insulation score** (IS) is defined for a bin as an average number of interactions that occur across this bin in some vicinity of the bin (Crane et al., 2015):

\[
\text{IS} = \frac{1}{k^2} \sum_{m \in M, n \in N} C(m, n),
\]

where \( N = \{n_1, n_2, \ldots, n_k\} \), \( M = \{m_1, m_2, \ldots, m_k\} \), \( C(m, n) \) is a number of interactions between bin \( m \) and bin \( n \).
We expect that IS has local minimums at TAD borders.

Lajoie et al., 2015, adapted
We expect that IS has local minimums at TAD borders.

IS plot is often called **insulation profile**.

Lajoie et al., 2015, adapted
Insulation score

IS can be calculated using a square window sliding along the diagonal of a contact matrix: average number of interactions in this window is the insulation score value (Crane et al., 2015, adapted):
Insulation score

IS can be calculated using a square window sliding along the diagonal of a contact matrix: average number of interactions in this window is the insulation score value (Crane et al., 2015, adapted):

![Insulation score diagram]
Insulation score

IS can be calculated using a square window sliding along the diagonal of a contact matrix: average number of interactions in this window is the insulation score value (Crane et al., 2015, adapted):
Insulation score

IS can be calculated using a square window sliding along the diagonal of a contact matrix: average number of interactions in this window is the insulation score value (Crane et al., 2015, adapted):
Insulation score

IS can be calculated using a square window sliding along the diagonal of a contact matrix: average number of interactions in this window is the insulation score value (Crane et al., 2015, adapted):
Insulation score

IS can be calculated using a square window sliding along the diagonal of a contact matrix: average number of interactions in this window is the insulation score value (Crane et al., 2015, adapted):
Insulation score

IS can be calculated using a square window sliding along the diagonal of a contact matrix: average number of interactions in this window is the insulation score value (Crane et al., 2015, adapted):
Insulation score

IS calculation scheme (Crane et al., 2015):

[Diagram showing the slide insulation square along diagonal with numbers and colors representing different insulation scores.]
Insulation score

TAD calling with IS (Crane et al., 2015):

- Calculate IS along a chromosome.
**Insulation score**

TAD calling with IS (Crane et al., 2015):

- Calculate IS along a chromosome.
- Normalize each IS value: \( IS := \log_2 \frac{IS}{IS_{avg}} \), where \( IS_{avg} \) is the mean of all IS values for the chromosome.
TAD calling with IS (Crane et al., 2015):

- Calculate IS along a chromosome.
- Normalize each IS value: $\text{IS} := \log_2 \frac{\text{IS}}{\text{IS}_{\text{avg}}}$, where IS$_{\text{avg}}$ is the mean of all IS values for the chromosome.
- Calculate $\Delta$ values for each bin $i$ (Crane et al., 2015, Extended Data):
Insulation score

TAD calling with IS (Crane et al., 2015):

- Calculate IS along a chromosome.
- Normalize each IS value: $IS := \log_2 \frac{IS}{IS_{avg}}$, where $IS_{avg}$ is the mean of all IS values for the chromosome.
- Calculate $\Delta$ values for each bin $i$. $\Delta_i = 0$ at all IS peaks and valleys (minimums) (Crane et al., 2015, adapted):
Insulation score

TAD calling with IS (Crane et al., 2015):
- Calculate IS along a chromosome.
- Normalize each IS value: \( IS := \log_2 \frac{IS}{IS_{avg}} \), where \( IS_{avg} \) is the mean of all IS values for the chromosome.
- Calculate \( \Delta \) values for each bin \( i \). \( \Delta_i = 0 \) at all IS peaks and valleys (minimums) (Crane et al., 2015, adapted):
  - TAD border is called at bin \( i \) if \( \Delta_i = 0 \), the nearest \( \Delta \) local max (\( \Delta_{max} \)) is to the left of bin \( i \), the nearest \( \Delta \) local min (\( \Delta_{min} \)) is to the right, and \( S_i \equiv \Delta_{max} - \Delta_{min} > 0.1 \). \( S_i \) is called border (boundary) strength.
  - TAD is called between two borders.

![Diagram of Local minima determination](image)


- **Crane et al., 2015** published their Perl script for TAD calling with IS.
- Crane et al., 2015 published their Perl script for TAD calling with IS.
- They called TAD borders with IS to see how they change in *C. elegans* X chromosome due to dosage compensation complex (DCC) depletion (Crane et al., 2015, adapted):
Crane et al., 2015 published their Perl script for TAD calling with IS. They called TAD borders with IS to see how they change in *C. elegans* X chromosome due to dosage compensation complex (DCC) depletion.

Barutcu et al., 2015 called TADs with IS to see differences in higher order chromatin structure between MCF-10A mammary epithelial and MCF-7 breast cancer cell lines.
Contrast index
Contrast index
Contrast index is defined as follows (Van Bortle et al., 2014, Alekseyenko et al., 2015):

$$CI = \frac{A + B}{C},$$

where $A$ is a total number of interactions to the left of bin $i$ in $L$-vicinity, $B$ is a total number of interactions to the right of bin $i$ in $L$-vicinity, and $C$ is a number of interactions that occur over bin $i$ from the left $L$-vicinity to the right.
Contrast index

CI calculation using a contact matrix (fig. is based on Crane et al., 2015):
Contrast index

CI calculation using a contact matrix (fig. is based on Crane et al., 2015):

\[ CI = \frac{\sum_A + \sum_B}{\sum_C}, \]

where \( \sum_A, \sum_B, \sum_C \) are sums of elements in A, B, and C contact submatrices, respectively.
TAD is called between two bins with CI values higher than some threshold.
TAD is called between two bins with CI values higher than some threshold.

No tool (script) was published for CI calculation.
Contrast index

- TAD is called between two bins with CI values higher than some threshold.
- No tool (script) was published for CI calculation.
- CI was used for TAD calling and TAD border strength assessment in several papers.
CI was used for TAD calling and TAD border strength assessment in several papers. E. g.:

- **Van Bortle et al., 2014** studied a relationship between TAD border strength and architectural proteins binding site (APBS) abundance (fig. is adapted):
Contrast index

- CI was used for TAD calling and TAD border strength assessment in several papers. E. g.:
  - Li et al., 2015 studied TAD border strength decline in *Drosophila* cells after heat-shock:
**Conclusion**

- **TADs** are stable and evolutionary conserved units of transcription regulation in mammals. Some similar self-interacting domains were found in other Eukaryotic species.
TADs are stable and evolutionary conserved units of transcription regulation in mammals. Some similar self-interacting domains were found in other Eukaryotic species.

Pros and cons of considered TAD calling methods:
Conclusion

- **TADs** are stable and evolutionary conserved units of transcription regulation in mammals. Some similar self-interacting domains were found in other Eukaryotic species.

- **Pros and cons** of **considered TAD calling methods**:
  - DI, IS, and CI are intuitive and inferred directly from TAD definition.
Conclusion

- **TADs** are stable and evolutionary conserved units of transcription regulation in mammals. Some similar self-interacting domains were found in other Eukaryotic species.

- **Pros** and **cons** of considered TAD calling methods:
  - DI, IS, and CI are intuitive and inferred directly from TAD definition.
  - They can be used both for TAD calling and TAD border strength assessment.
**Conclusion**

- **TADs** are stable and evolutionary conserved units of transcription regulation in mammals. Some similar self-interacting domains were found in other Eukaryotic species.

- **Pros** and **cons** of considered TAD calling methods:
  - DI, IS, and CI are intuitive and inferred directly from TAD definition.
  - They can be used both for TAD calling and TAD border strength assessment.
  - DI, IS, and CI are easy to compute: each of them can be calculated in $O(NK)$ time for one chromosome, where $N$ is a number of bins in a chromosome, and $2K$ is a number of bins in the 2$L$-vicinity of each bin. Typically, $K$ is much less than $N$. 
**Conclusion**

- **TADs** are stable and evolutionary conserved units of transcription regulation in mammals. Some similar self-interacting domains were found in other Eukaryotic species.

- **Pros** and **cons** of considered TAD calling methods:
  - DI, IS, and CI are intuitive and inferred directly from TAD definition.
  - They can be used both for TAD calling and TAD border strength assessment.
  - DI, IS, and CI are easy to compute: each of them can be calculated in $O(NK)$ time for one chromosome, where $N$ is a number of bins in a chromosome, and $2K$ is a number of bins in the $2L$-vicinity of each bin. Typically, $K$ is much less than $N$.
  - We need an arbitrary threshold / percentile or a kind of HMM to call TADs with these metrics.
**Conclusion**

- **TADs** are stable and evolutionary conserved units of transcription regulation in mammals. Some similar self-interacting domains were found in other Eukaryotic species.

- **Pros and cons of considered TAD calling methods:**
  - DI, IS, and CI are intuitive and inferred directly from TAD definition.
  - They can be used both for TAD calling and TAD border strength assessment.
  - DI, IS, and CI are easy to compute: each of them can be calculated in $O(NK)$ time for one chromosome, where $N$ is a number of bins in a chromosome, and $2K$ is a number of bins in the $2L$-vicinity of each bin. Typically, $K$ is much less than $N$.
  - We need an arbitrary threshold / percentile or a kind of HMM to call TADs with these metrics.
  - There are almost no published and well-tested tools for TAD calling using these metrics.
Conclusion

- **TADs** are stable and evolutionary conserved units of transcription regulation in mammals. Some similar self-interacting domains were found in other Eukaryotic species.

- **Pros and cons** of considered TAD calling methods:
  - DI, IS, and CI are intuitive and inferred directly from TAD definition.
  - They can be used both for TAD calling and TAD border strength assessment.
  - DI, IS, and CI are easy to compute: each of them can be calculated in $O(NK)$ time for one chromosome, where $N$ is a number of bins in a chromosome, and $2K$ is a number of bins in the $2L$-vicinity of each bin. Typically, $K$ is much less than $N$.
  - We need an arbitrary threshold / percentile or a kind of HMM to call TADs with these metrics.
  - There are almost no published and well-tested tools for TAD calling using these metrics.
  - DI, IS, and CI can’t enable us to call a TAD hierarchy (a TAD with its sub-TADs) as a whole.
Conclusion

**Pros** and **cons** of considered methods:

- DI, IS, and CI are intuitive and inferred directly from TAD definition.
- They can be used both for TAD calling and TAD border strength assessment.
- DI, IS, and CI are easy to compute: each of them can be calculated in $O(NK)$ time for one chromosome, where $N$ is a number of bins in a chromosome, and $2K$ is a number of bins in the $2L$-vicinity of each bin. Typically, $K$ is much less than $N$.
- We need an arbitrary threshold / percentile or a kind of HMM to call TADs with these metrics.
- There are almost no published and well-tested tools for TAD calling using these metrics.
- DI, IS, and CI can’t enable us to call a TAD hierarchy (a TAD with its sub-TADs) as a whole.

**In Part 2** I’ll consider *some* of the following much more complicated methods and tools for TAD calling: Sexton et al., 2012; Hou et al., 2012; Armatus, 2014; HiCseg, 2014; Arrowhead algorithm, 2014; TADtree, 2015; TADBite.
1. Introduction

2. Topologically associating domains

3. TAD calling methods

4. Conclusion

5. Selected literature
Chromatin conformation overviews

Self-interacting chromatin domains in various species

- **Chromatin interaction domains (CIDs) in bacterium Caulobacter crescentus:** Le T. B. et al. 2013. High-resolution mapping of the spatial organization of a bacterial chromosome *Science* 342(6159): 731–734.


Chromatin conformation capture methods:


- **Some Hi-C derivatives:**
Hi-C data processing and analysis

Overviews:


Hi-C data correction:

TAD calling methods

Covered in this overview:


Thank you!

Sam Rose. Epigenetics and organisation