Non-deformable Sets of 3C Constraints

Carl Kingsford
Assistant Professor
Center for Bioinformatics & Computational Biology
Department of Computer Science

Joint work with Geet Duggal, Hao Wang, Cara Treglio, Michelle Girvan, and Sridhar Hannenhalli
The Problem

- How confident are we in this structure?
- Is there a unique embedding consistent with the constraints?
- Which regions have unique embeddings?
- Which interactions are important for creating unique embeddings?
The Problem

* How confident are we in this structure?

* Is there a unique embedding consistent with the constraints?

* Which regions have unique embeddings?

* Which interactions are important for creating unique embeddings?
Multiple Solutions → Uncertainty

- RMSD between 100 different structures for Bau et al. segment of human chromosome 16
- Duan et al. optimization with random initial conditions
- red = high RMSD, blue = low RMSD
- See also: Rousseau et al., BMC Bioinformatics, 2011
Rigidity (Bar-Joint)

A graph is rigid in \( d \) dimensions if: when it is embedded in \( \mathbb{R}^d \), there is no continuous motion of the vertices that preserves edge lengths (except translations and rotations).

Technicalities:
- doesn’t imply unique embedding: there may be multiple embeddings, but “you can get there from here.” There are \( < \infty \).
- requires embeddings to be algebraically independent (generalization of general position). But: a random embedding is algebraically independent with probability 1.

Thm. If one embedding is rigid, then they all are.

⇒ rigidity is a property of the graph, not an embedding (or any set of distances)

(Gluck, 1975)
Example Rigid Components (3D)

- Joint (universal ball)
- Bar

- "Nunchuk": Floppy
- Triangle: Rigid

- Hinge
- Rigid components

- Double Banana: Two Components
- "Rigidified" Double Banana

- Triple Banana: Four Components
Example Rigid Components (3D)

Joint (universal ball)

"Nunchuk": Floppy

Triangle: Rigid

Hinge

Rigid components

Double Banana: Two Components

"Rigidified" Double Banana

Bar

Triple Banana: Four Components
Augmented conformation graph: includes edges between adjacent DNA fragments

Edges correspond to types of distance constraints that are available during an embedding optimization.
Chromosome Conformation Graph:

Augmented conformation graph: includes edges between adjacent DNA fragments

Edges correspond to types of distance constraints that are available during an embedding optimization.
# Recent 3C Data Sets

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Genome</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lieberman-Aiden et al.</td>
<td>Human</td>
<td>100,1000</td>
</tr>
<tr>
<td>Duan et al.</td>
<td>Budding yeast</td>
<td>F,10</td>
</tr>
<tr>
<td>Tanizawa et al.</td>
<td>Fission yeast</td>
<td>20</td>
</tr>
<tr>
<td>Bau et al.</td>
<td>Human chr. 16</td>
<td>F</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment</th>
<th># Vertices</th>
<th>Maximum intra-chromosomal frequency</th>
<th>Maximum inter-chromosomal frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lieberman-Aiden et al. GM06690</td>
<td>2,882</td>
<td>29,931</td>
<td>6,068</td>
</tr>
<tr>
<td>Lieberman-Aiden et al. K562</td>
<td>2,882</td>
<td>41,124</td>
<td>3,331</td>
</tr>
<tr>
<td>Duan et al.</td>
<td>4,193</td>
<td>4,683</td>
<td>107</td>
</tr>
<tr>
<td>Tanizawa et al.</td>
<td>619</td>
<td>35.25</td>
<td>13.75</td>
</tr>
<tr>
<td>Bau et al. GM12878</td>
<td>55</td>
<td>5,823</td>
<td>-</td>
</tr>
<tr>
<td>Bau et al. K562</td>
<td>55</td>
<td>13,686</td>
<td>-</td>
</tr>
</tbody>
</table>
Finding Rigid Components

- Testing rigidity of entire graph: randomly embed & check rank of rigidity matrix
- “2d pebble game” is exact combinatorial algorithm for rigid components in 2D (Jacobs & Hendrickson, 1997)
- No known polynomial-time algorithm for 3D

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Yeast ChrII (246 nodes, 2148 edges)</th>
<th>Entire Yeast Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relaxation (Chubynsky &amp; Thorpe, 2008)</td>
<td>days</td>
<td>heat death of universe?</td>
</tr>
<tr>
<td>Pebble Game 3D</td>
<td>&lt; 10 sec (but wrong)</td>
<td>&lt; 2 mins (but wrong)</td>
</tr>
<tr>
<td>Ours</td>
<td>&lt; 10 sec</td>
<td>a few hours (variant: a few minutes)</td>
</tr>
</tbody>
</table>
Our Method for Finding Rigid Components in 3D:

- Merge components that overlap by $\geq 3$ vertices

Diagram:

- 3D Pebble Game
  - Rigidity Rank Test
  - Rigidity Rank Test
  - Rigidity Rank Test
  - Rigidity Rank Test

- Greedy Growing
  - Convert to Body-Bar-and-Hinge rigidity problem & solve

- Generic 3-Gluing Lemma (Whiteley, 1996)
  - (Hendrickson, 1992)
Greedy Growing

Start with a triangle that:

1. doesn't overlap any previously-chosen triangle, and
2. has the largest # of connections to unused vertices

Repeatedly add a vertex that:

1. has $\geq 3$ connections to the current set of vertices, and
2. has the largest # of connections to unused vertices

Stop when no such vertices exist

Repeat all the above until you can’t find any more non-overlapping triangles

Vertex 3-Addition
Lemma (Whiteley, 1996)
Reduction to Body-Bar-and-Hinge

Treat rigid subgraphs found so far as rigid bodies with “ports”

Bodies can be connected by bars (that don’t share a port)

Bodies that share 2 ports are connected by a hinge (hinges can’t overlap)

Bodies that share 1 port are not allowed

Identifying rigid components in BBH frameworks is solvable in $O(n^2)$ time (Lee et al., 2005)
Large Rigid Components

* Including all edges $\rightarrow$ completely rigid

* Even throwing away the lowest-frequency 98.8% of the interactions $\rightarrow$ nearly completely rigid:

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Unaugmented</th>
<th>Augmented</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>graph size</td>
<td>rigid component</td>
</tr>
<tr>
<td>GM06690</td>
<td>2,880</td>
<td>2,879</td>
</tr>
<tr>
<td>K562</td>
<td>2,874</td>
<td>2,874</td>
</tr>
<tr>
<td>Budding yeast</td>
<td>3,172</td>
<td>2,880</td>
</tr>
<tr>
<td>Fission yeast</td>
<td>611</td>
<td>590</td>
</tr>
</tbody>
</table>

* Only minor differences between augmented vs. unaugmented
Components naturally break apart after throwing away more interactions.
Rigid components on Duan et al.'s model
(99.6% interactions removed)
Rigid components on Tanizawa et al.’s model (99.0% interactions removed)
Removing Short Interactions

*Removing interactions spanning ≤ 75kb → destruction of nearly all rigid components*
Chromosome Rigidity

* Rigidity useful pre-processing step to identify which regions of structure are most believable

* Raw 3C data is way more than enough to fix a finite # of possible embeddings

* Filtered 3C data is too (up to some high fraction of edges removed)

* Short interactions generally crucial for rigidity

* Unaugmented vs. augmented doesn’t make much difference

* Pebble game identifies redundant edges and extracts sets of 3n-6 edges needed for rigidity → faster embedding & multiple solutions
Rigidity Caveats

* Assumes distance constraints are exact (corresponding theory of inexact constraints not as well developed)

* Rigid $\neq$ unique (counting # of possible embeddings is another hard problem)

* Doesn’t consider consistency of distances
Spatial Enrichment – Problems

(1) Large sets $\rightarrow$ small P-values

Set of interactions

set of interest
(e.g. co-regulated genes)
size = $s$

Sample $s$ items and take the average

$\mathcal{N}(\mu, \sigma)$

$\mathcal{N}(\mu, \frac{\sigma}{\sqrt{s}})$

(2) Density is imperfect measure of proximity:

A & B are close by transitivity

Weak evidence that A & B are close
Spatial Enrichment – Problems

(1) Large sets → small P-values

Set of interactions

set of interest
(e.g. co-regulated genes)

size = \( s \)

Sample \( s \) items and take the average

\( \mathcal{N}(\mu, \sigma) \)

\( \mathcal{N}(\mu, \frac{\sigma}{\sqrt{s}}) \)

Correction by fitting a parametric distribution

(2) Density is imperfect measure of proximity:

A & B are close by transitivity

Richer topological proximity measures (shortest path, maximum flow, densest subgraph)
Thanks!

Joint work with
  Geet Duggal (rigidity)
  Hao Wang (statistics)
  Cara Treglio (multiple embeddings)
  Michelle Girvan
  Sridhar Hannenhalli

Thanks to PSB & the NIH for a travel award.

Partially supported by:
  National Science Foundation [CCF-1053918, EF-0849899, and IIS-0812111]
  National Institutes of Health [1R21AI085376]
  University of Maryland Institute for Advanced Studies New Frontiers Award