Folding and Unfolding in Graphs and DNA

Tero Harju

Work done together with
A. Ehrenfeucht, I. Petre, and G. Rozenberg
Folding / unfolding

- **Folding and unfolding** problems are popular in many pure and applied areas:
  - **Protein folding**
  - **Map folding**: given creases can you fold the map
Life is Hard (for a geometrician)

Shephard 1975: Is it true that every (convex) polyhedra can be cut along edges and unfolded to form a simple polygon in the plane?

The inverse problem: Characterize polygons that (re)fold to polyhedra.
There are concave polygons that cannot be so folded.
Can you have a sequence of foldings of a square paper so that you obtain larger perimeter than the original one?
Here: a graph theoretic model for the splicing process motivated by recombination in ciliates.
Here: a graph theoretic model for the splicing process motivated by recombination in ciliates.

- circular DNA molecule $\leftrightarrow$ recombined molecule
- Cycle $\leftrightarrow$ folding $\leftrightarrow$ unfolding
Consider graphs having possibly **multiple edges and loops**, with **edge colouring** and **edge labelling**.

Each edge is **oriented** in the opposite directions:
Consider graphs having possibly multiple edges and loops, with edge colouring and edge labelling.

Each edge is oriented in the opposite directions:

In general: the colours are 1, 2, ..., k.

Bicolouring with green and blue edges corresponds to MDS and IES regions in micronuclear genome of ciliates.
With(out) labels

Labels are complementary strings: DNA double strands

\[
\begin{align*}
\text{AGTC} & = \text{GACT} \\
\text{TCAG} & = \text{CTGA}
\end{align*}
\]

string \( s \)

inverse \( \bar{s} \)
With(out) labels

Labels are complementary strings: DNA double strands

\[
\begin{align*}
\text{AGTC} & = \text{GACT} \\
\text{TCAG} & = \text{CTGA}
\end{align*}
\]

For simplicity, we usually suppress the labels.

However, they are important for the directions.
With(out) labels

Labels are complementary strings: DNA double strands

\[
\begin{align*}
\overline{AGTC} & = GACT \\
TCAG & = CTGA
\end{align*}
\]

string \( s \)

inverse \( \bar{s} \)

Draw only \( e \) or \( \bar{e} \).
Valency

- **valency** $\text{val}(x)$: the number of edges entering $x$.

- The $c$-valency:

$$\text{val}(x, c) = \left| \{e = (y, x) \in E \mid e \text{ has colour } c \} \right|$$
Valency

- **valency** \( \text{val}(x) \): the number of edges entering \( x \).

- The **\( c \)-valency**:
  \[
  \text{val}(x, c) = |\{e = (y, x) \in E \mid e \text{ has colour } c\}|
  \]

- **Balanced** at a vertex \( x \): for each colour \( c \in [1, k] \),
  \[
  \text{val}(x, c) \leq \frac{\text{val}(x)}{2}
  \]
  For **bicoloured**: equality

- The graph is **balanced**, if balanced at each vertex.
Example

- **bicoloured**
- **balanced**
- **val**(*x, c*) is 1 or 2 for each *x*. 
Genes

All edges of certain colour(s) are to be spliced together
in predetermined order and direction
(determined by pointers = pairs of vertices)
to form contiguous monochromatic paths representing assembled genes.
**Kotzig’s theorem**

- An alternating Euler tour in $\gamma$:
  - a closed walk visiting each edge exactly once
  - consecutive edges have different colours.
- $\gamma$ is even, if all valencies are even.

**THEOREM (Kotzig)**

A graph has an alternating Euler tour if and only if it is connected, even and balanced.
Reorientations

Exchange operation transforms closed walks:

\[ \pi_1 \pi_2 \pi_3 \pi_4 \pi_5 \rightarrow \pi_1 \pi_4 \pi_3 \pi_2 \pi_5 \]

where \( \pi_2 \) and \( \pi_4 \) have the same ends.
Reorientations

- **Exchange operation** transforms closed walks:
  \[ \pi_1 \pi_2 \pi_3 \pi_4 \pi_5 \rightarrow \pi_1 \pi_4 \pi_3 \pi_2 \pi_5 \]
  where \( \pi_2 \) and \( \pi_4 \) have the same ends.

- **Reflection operation** transforms closed walks:
  \[ \pi_1 \pi_2 \pi_3 \rightarrow \pi_1 \pi_2 \pi_3 \]
  where \( \pi_2 \) is a closed walk.
THEOREM (Pevzner)

Let $\gamma$ be bicoloured with alternating Euler tours $T_1$ and $T_2$. Then $T_1$ is obtained from $T_2$ by applying exchange and reflection operations preserving alternating closed walks.
Pairing functions in even graphs $\gamma$

A pairing $\psi_x$ of a vertex $x$ is a bijection on edges:

\[
\begin{align*}
\text{incoming} & \iff \text{outgoing} \\
\begin{array}{c}
\uparrow \\
\text{e}
\end{array} & \leftrightarrow \\
\begin{array}{c}
\downarrow \\
\psi_x(e)
\end{array}
\end{align*}
\]

respect inverses: $\psi_x(\psi_x(e)) = \overline{e}$

and so also loops: $\psi_x(e) = \overline{e} \iff e = (x, x)$
Pairing functions in even graphs $\gamma$

- A pairing $\psi_x$ of a vertex $x$ is a bijection on edges:

\[
\begin{align*}
\text{incoming} & \leftrightarrow \text{outgoing} \\
\psi_x(e) & \quad \psi_x(e) \\
\end{align*}
\]

- respect inverses: $\psi_x(\psi_x(e)) = \overline{e}$
- and so also loops: $\psi_x(e) = \overline{e} \iff e = (x, x)$

- A pairing function $\psi$ of the full graph $\psi: x \mapsto \psi_x$:
  $\psi_x$ is a pairing for each vertex $x$. 

\[ \text{diagram} \]
Tucker’s theorem

For each \( e_1 \in E \), let \( \pi_\psi(e_1) = e_1 e_2 \ldots e_k \) with \( e_i = (x_i, x_{i+1}) \) be a maximal trail such that \( \psi_{x_i}(e_{i-1}) = e_i \) for all \( i \).
Tucker’s theorem

For each $e_1 \in E$, let $\pi_{\psi}(e_1) = e_1e_2 \ldots e_k$ with $e_i = (x_i, x_{i+1})$ be a maximal trail such that $\psi_{x_i}(e_{i-1}) = e_i$ for all $i$.

THEOREM (Tucker)

Maximal trails $\pi_{\psi}(e)$ are well defined and closed. The edge sets of the maximal trails partition $E$. 
Recombination graphs

In gene assembly we are interested in recombination graphs:

- **bicoloured**
- \( \text{val}(x) = 2 \) or \( 4 \) for each \( x \).
- Balanced at vertices of valency \( \text{val}(x) = 4 \).

- **Green** edges make a gene
- **Blue** edges are
  - IESs or introns or ...
Pairing recombination graphs

Each recombination graph $\gamma$ has natural pairing function:

- $\psi_x$ matches edges with the same colour if $\text{val}(x) = 4$
- the only possible edges for $\text{val}(x) = 2$. 

\begin{itemize}
  \item \begin{tikzpicture}
    \path[每边宽度=5pt, 每边颜色=red]
    (0,0) edge (0.5,0.5)
    (0,0) edge (0.5,-0.5)
    (0.5,0) edge (1,0.5)
    (0.5,0) edge (1,-0.5)
  \end{tikzpicture}
  \begin{tikzpicture}
    \path[每边宽度=5pt, 每边颜色=green]
    (0,0) edge (0.5,0.5)
    (0,0) edge (0.5,-0.5)
    (0.5,0) edge (1,0.5)
    (0.5,0) edge (1,-0.5)
  \end{tikzpicture}
  \begin{tikzpicture}
    \path[每边宽度=5pt, 每边颜色=blue]
    (0,0) edge (0.5,0.5)
    (0,0) edge (0.5,-0.5)
    (0.5,0) edge (1,0.5)
    (0.5,0) edge (1,-0.5)
  \end{tikzpicture}
\end{itemize}
Graph theoretical framework

The process of gene assembly is divided to two stages:

- **Folding** circular graphs (circular DNA).
  - Folding is determined by **pointer sets**.
  - A pair \( p = \{x, y\} \ (x \neq y) \) is a **pointer**.
  - A set \( P \) of **disjoint** pointers is a **pointer set**.
Graph theoretical framework

The process of gene assembly is divided to two stages:

- **Folding** circular graphs (circular DNA).
  Folding is determined by pointer sets.
  - A pair $p = \{x, y\} \ (x \neq y)$ is a pointer.
  - A set $P$ of disjoint pointers is a pointer set.

- **Unfolding** graphs by splitting vertices.
  Unfolding is determined by a pairing function $\psi$. 

Actin I gene in *O.nova*

The parts of the gene are numbered: 1, 2, ..., 9. The ends of these are enumerated around the cycle.
Actin I gene in O.nova

The pointers $p_1, \ldots, p_8$ are:

$\{13, 1\}$, $\{2, 3\}$, $\{4, 7\}$, $\{8, 5\}$,
$\{6, 9\}$, $\{10, 17\}$, $\{18, 11\}$, $\{16, 14\}$
Let $p = \{x, y\}$ be a pointer.

The $p$-folded graph $\gamma \ast p$ on $V' = (V \setminus \{x, y\}) \cup \{p\}$

is obtained by identifying the elements of $p$. 

\[\begin{array}{cc}
\text{a} & \text{x} & \text{b} \\
\text{c} & \text{y} & \text{d} \\
\end{array}\quad \begin{array}{cc}
\text{a} & \text{p} & \text{b} \\
\text{c} & \text{d} \\
\end{array}\]
Remark

The $p$-folded graph $\gamma \ast p$ has

- the same set of edges (with different ends),
- the same labelling and colouring as $\gamma$.
- But $|V_{\gamma \ast p}| = |V_{\gamma}| - 1$.
- Edges may become loops or parallel in $\gamma \ast p$. 

\[\text{Diagram:}\]

- Graph $\gamma$ with nodes $a$, $x$, $y$, and $d$.
- Graph $\gamma \ast p$ with an additional node $p$ and edges connecting it to other nodes.
Commutation

LEMMA

Let \( \{p, q\} \) be a pointer set of \( \gamma \). Then
\[
(\gamma * p) * q = (\gamma * q) * p.
\]
**LEMMA**

Let \( \{p, q\} \) be a pointer set of \( \gamma \). Then
\[
(\gamma * p) * q = (\gamma * q) * p.
\]

**THEOREM**

For a pointer set \( P = \{p_1, p_2, \ldots, p_m\} \), the \( P \)-folded graph:
\[
\gamma * P = \gamma * p_1 * \ldots * p_m
\]

If \( P_1, P_2, \ldots, P_n \) are disjoint pointer sets, then
\[
\gamma * P_1 * P_2 * \ldots * P_n = \gamma * \bigcup_{i=1}^{n} P_i
\]
Again Actin I gene

The pointers $p_1, \ldots, p_8$:

$$\{13, 1\}, \{2, 3\}, \{4, 7\}, \{8, 5\},$$

$$\{6, 9\}, \{10, 17\}, \{18, 11\}, \{16, 14\}$$
The folded actin I gene in O.nova
Let $\gamma$ be an even graph with a pairing function $\psi: x \mapsto \psi_x$. For a vertex $x$, let $\psi_x(e_{i1}) = \overline{e}_{i2}$, $e_{11}, e_{12}, e_{21}, e_{22}, \ldots, e_{m1}, e_{m2}$.
Let $x^1, x^2, \ldots, x^m$ be new vertices.

The $\psi$-unfolded graph at $x$ is obtained by redirecting w.r.t. $\psi_x(e_{i1}) = \bar{e}_{i2}$:

\[ \gamma \diamond \psi \ x \]
Remark

The $\psi$-unfolded graph has

- the same set of edges (with different ends),
- the same labelling and colouring as $\gamma$.
- Also, $\psi$ remains as a pairing function of $\gamma \diamond \psi \ x$. 
LEMMA
Let $\gamma$ be an even graph with a pairing function $\psi$. For $x \neq y$, $(\gamma \diamond_{\psi} x) \diamond_{\psi} y = (\gamma \diamond_{\psi} y) \diamond_{\psi} x$.

This allows us to define for $A = \{x_1, \ldots, x_m\}$:

$$\gamma \diamond_{\psi} A = \gamma \diamond_{\psi} x_1 \diamond_{\psi} x_2 \diamond_{\psi} \ldots \diamond_{\psi} x_m$$
The unfolded graph

Let \( F(\gamma) = \{ x \in V \mid \text{val}(x) \geq 4 \} \)

\( \gamma \diamond \psi F(\gamma) \) is the \( \psi \)-unfolded graph of \( \gamma \).

**LEMMA**

If \( \gamma \) is even with a pairing function \( \psi \), then \( \gamma \diamond \psi F(\gamma) \) is a disjoint union of cyclic graphs.
Fold-and-unfold

Let $\gamma$ be a graph with a pointer set $P$, $\psi$ a pairing function of the $P$-folded graph $\gamma \ast P$. Set

$$\gamma \otimes_\psi P = (\gamma \ast P) \diamond_\psi P$$

**LEMMA**

If $\gamma$ is a disjoint union of cycles, then so is $\gamma \otimes_\psi P$. 

- p. 30/38
Example

$P: 1s$ and $2s$

$(\gamma \ast P) \diamond F(\gamma)$
Assembled graphs of genomes

Let $\gamma$ be a bicoloured cyclic graph.

- **boundary vertex**: changes colour at $x$

- **(Gene) segment**: maximally 1-monochromatic path.
  So that its ends are boundary vertices.
A pair $G = (\gamma, P)$ is a genome, if

- $\gamma$ is a disjoint union of bicoloured cyclic graphs,
- $P$ is a pointer set of pairs of boundary vertices.

The assembled version of $G$ is $A(G) = \gamma \ast P$ (w.r.t. the natural pairing function).

Genomes $G$ and $G'$ are equivalent, if $A(G) = A(G')$. 

Genomes
Genes

- A segment of $\gamma \circ P$ is a noncircular gene of $\mathcal{G}$.

- Each 1-monochromatic cyclic component of $\gamma \circ P$ is a circular gene of $\mathcal{G}$.

- Genes are assembled from various (gene-) segments which may lie on different cycles of $\mathcal{G}$. 

\[ g_1 \quad g_4 \quad g_2 \quad g_3 \]

\[ g_1 g_2 g_3 g_4 \]
Final genome of actin I gene in O.nova
Invariance

Every **assembly strategy** produces the same **assembled genome**:

**THEOREM** (Invariance theorem)

For every partition $P_1, \ldots, P_m$ of $P$,

$$\gamma \odot P = \gamma \odot P_1 \odot P_2 \odot \ldots \odot P_m.$$
Intracyclic unfolding

An assembly strategy $P_1, \ldots, P_m$ of $G = (\gamma, P)$ is intracyclic, if genes are not produced from disjoint parts:

NOT:
**THEOREM**

Let $\mathcal{G}$ be a genome. There exists an **intracyclic assembly strategy** $P_1, P_2, \ldots, P_m$ with $1 \leq |P_i| \leq 2$ for all $i$.

**THEOREM**

Let $A(\mathcal{G}')$ be connected. There is a equivalent $\mathcal{G} = (\gamma, P)$ with **connected** $\gamma$ and an assembly strategy $P_1, P_2, \ldots, P_m$ such that $1 \leq |P_i| \leq 2$ for all $i$, and each intermediate $\gamma \, \ast \, \bigcup_{i=1}^{j} P_i$ is connected.