Topology of RNA and DNA

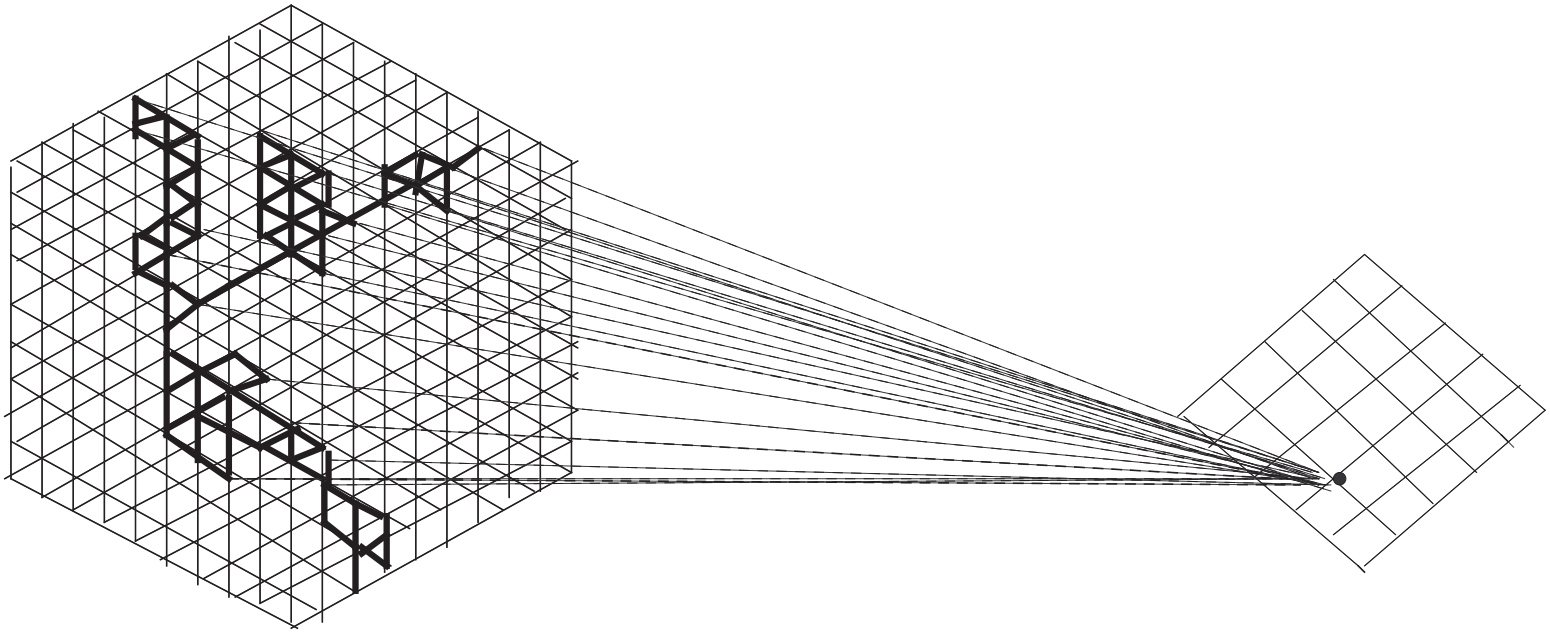
Christian M. Reidys

[1]University of Southern Denmark

September, 2013

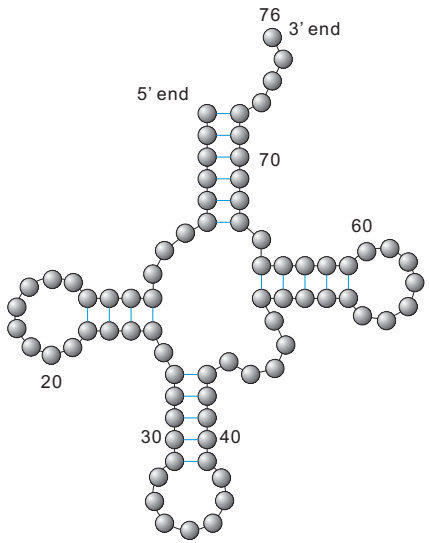
Outline

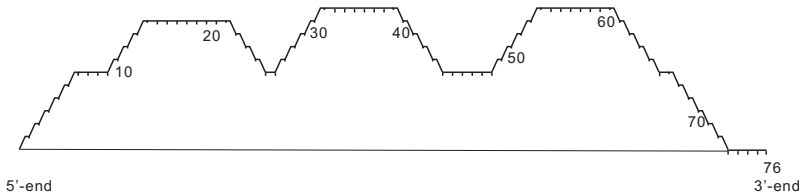
Neutral networks

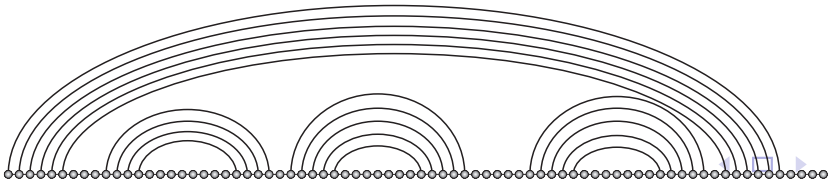


Reidys, C.M., Stadler, P.F., Schuster, P.K., 1997, *Generic Properties of Combinatory Maps and Neutral Networks of RNA Secondary Structures*, Bull. Math. Biol., **(59)**, 339-397

Secondary structures







Motzkin-path bijection

We translate a diagram without crossing to a path, from $(0, 0)$, reading from left to right:

- start point of an arc $\longrightarrow$ up-step
- end point of an arc $\longrightarrow$ down-step
- isolated vertex $\longrightarrow$ horizontal-step

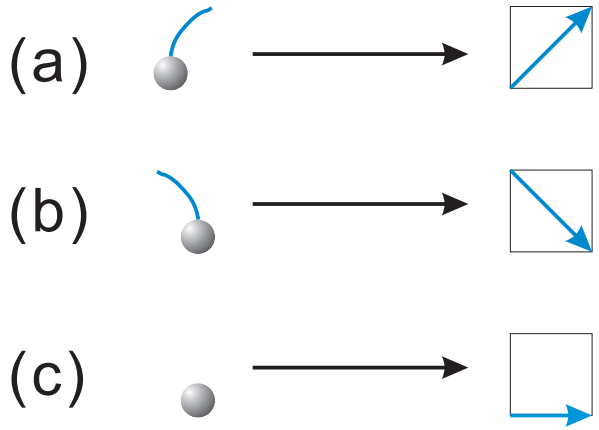


Figure: (a) An up-step when we meet a start point of an arc. (b) A down-step when we meet an end point of an arc. (c) A horizontal-step when we meet an isolated point.

Example

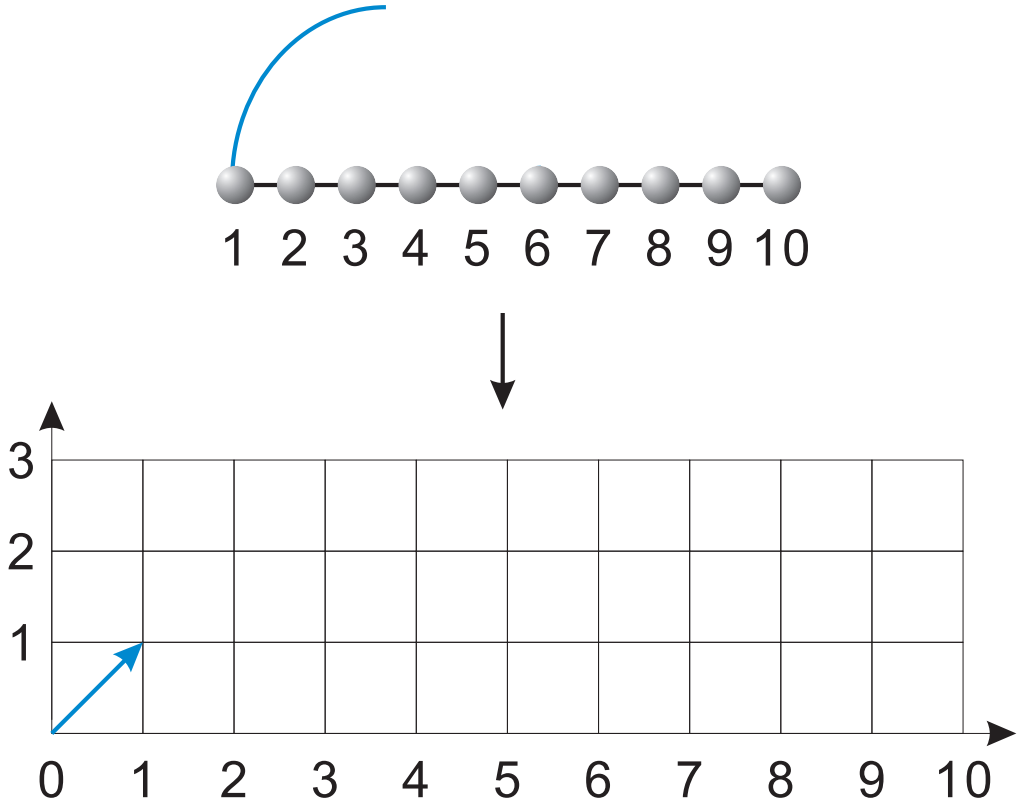


Figure: From a diagram to a path.

From a diagram to a path

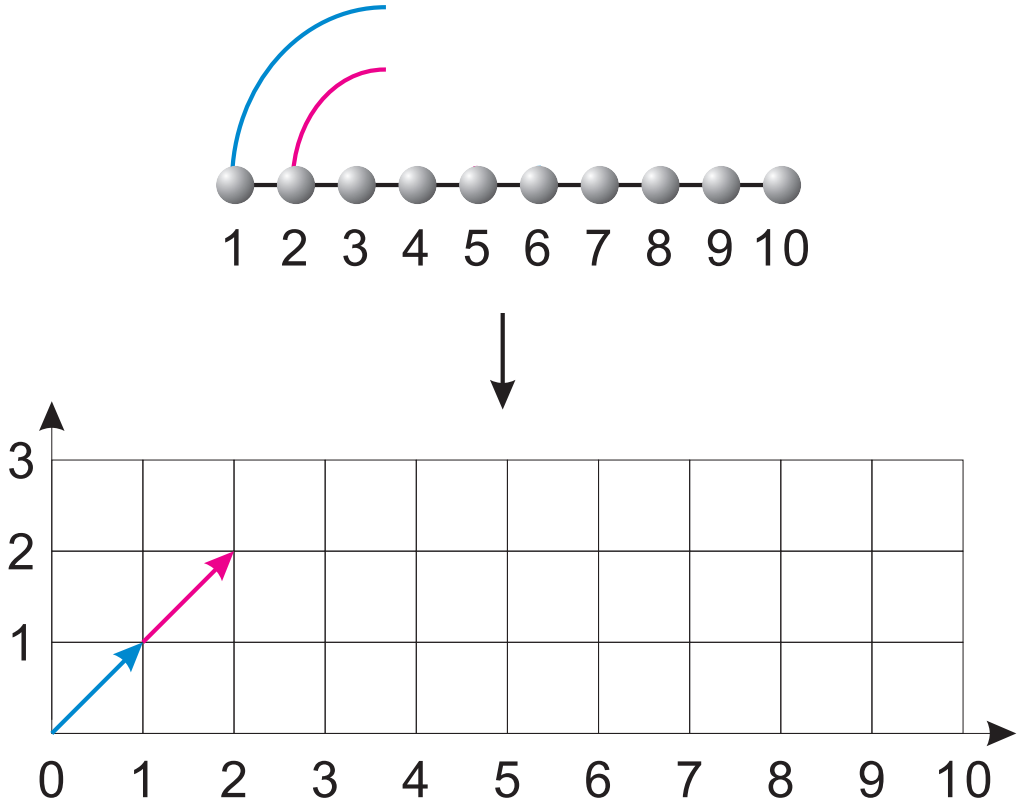


Figure: From a diagram to a path.

From a diagram to a path

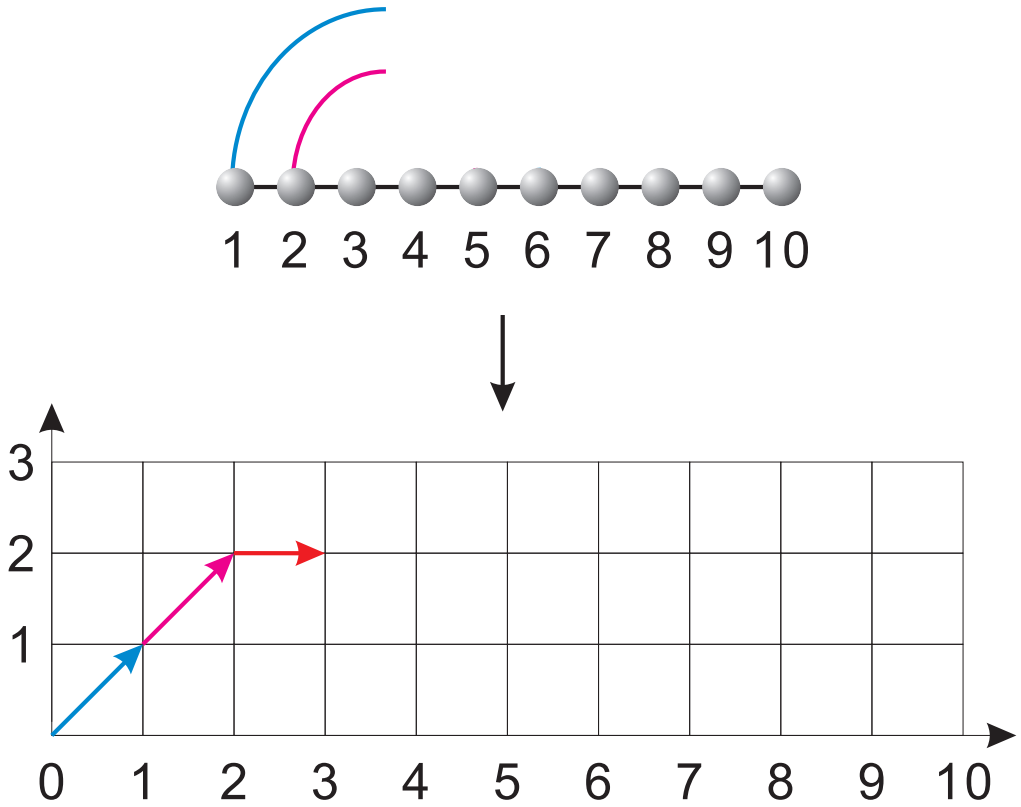


Figure: From a diagram to a path.

From a diagram to a path

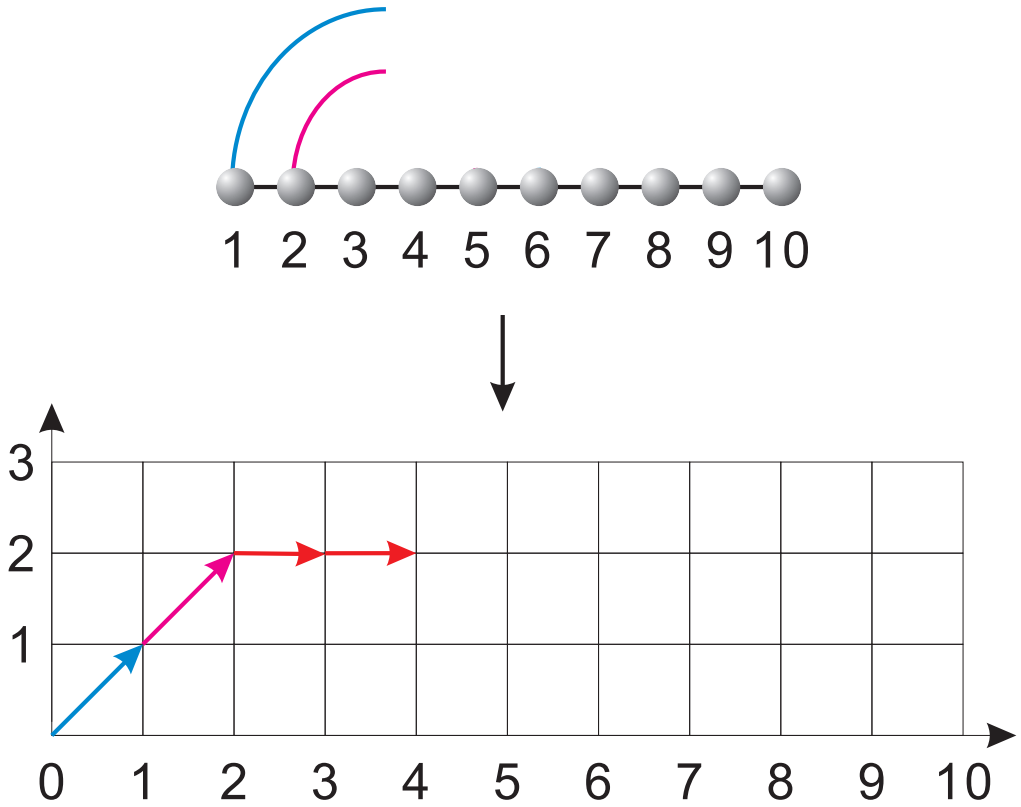


Figure: From a diagram to a path.

From a diagram to a path

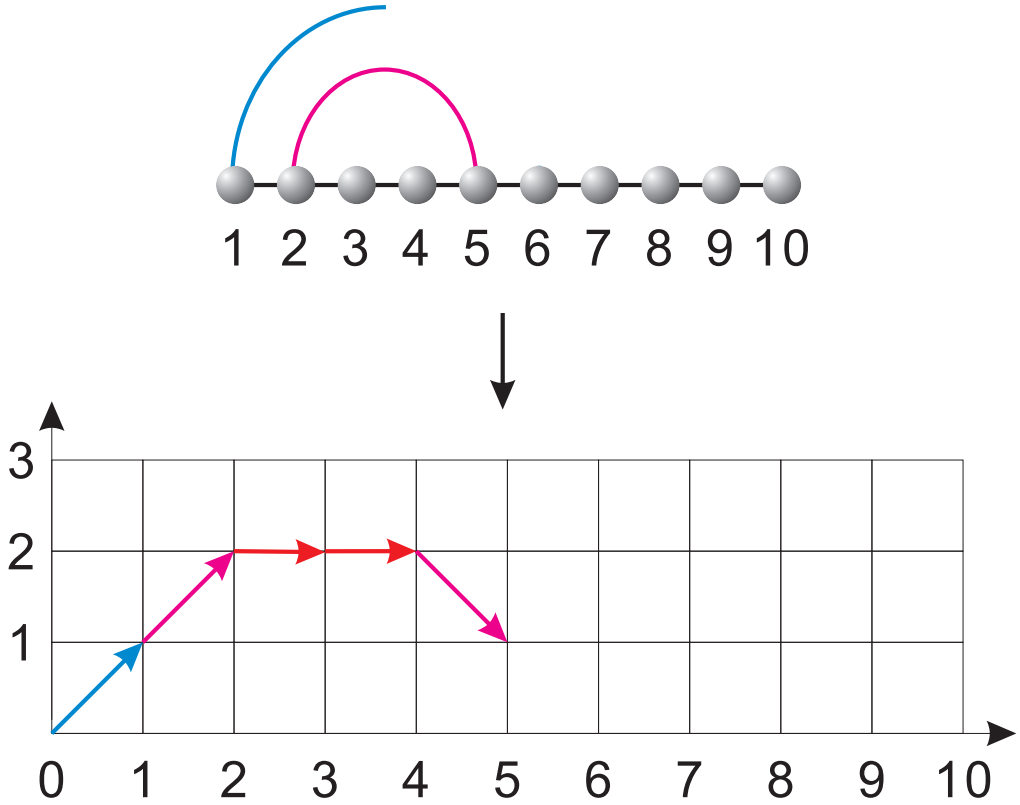


Figure: From a diagram to a path.

From a diagram to a path

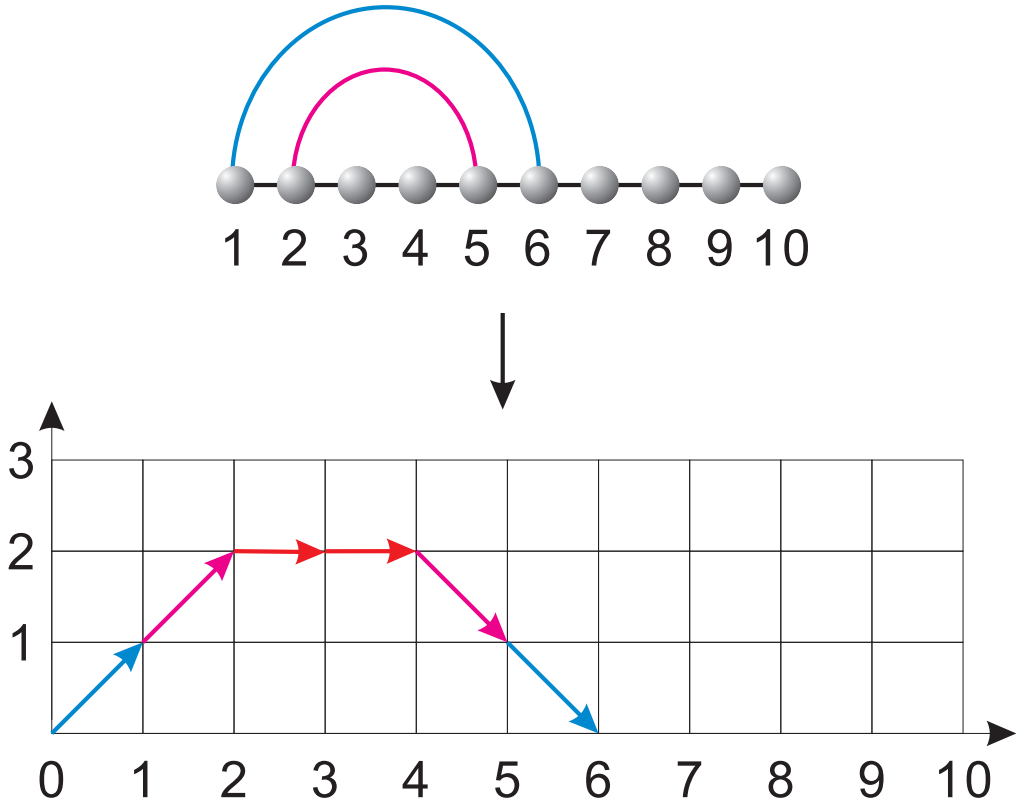


Figure: From a diagram to a path.

From a diagram to a path

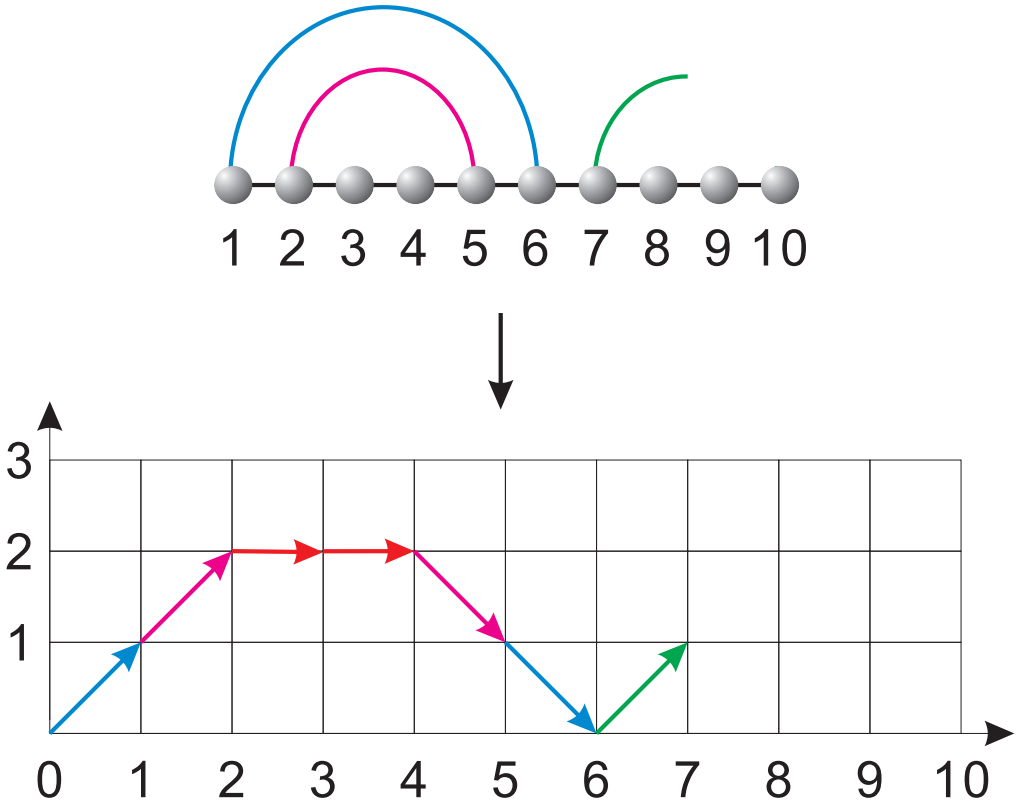


Figure: From a diagram to a path.

From a diagram to a path

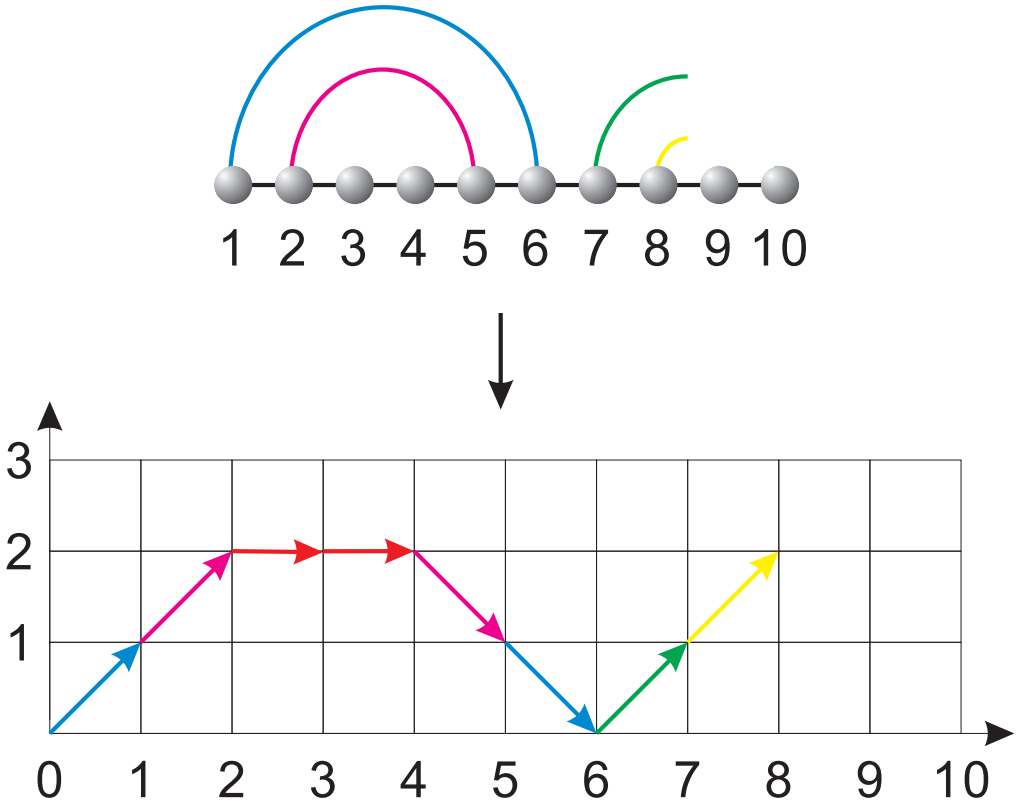


Figure: From a diagram to a path.

From a diagram to a path

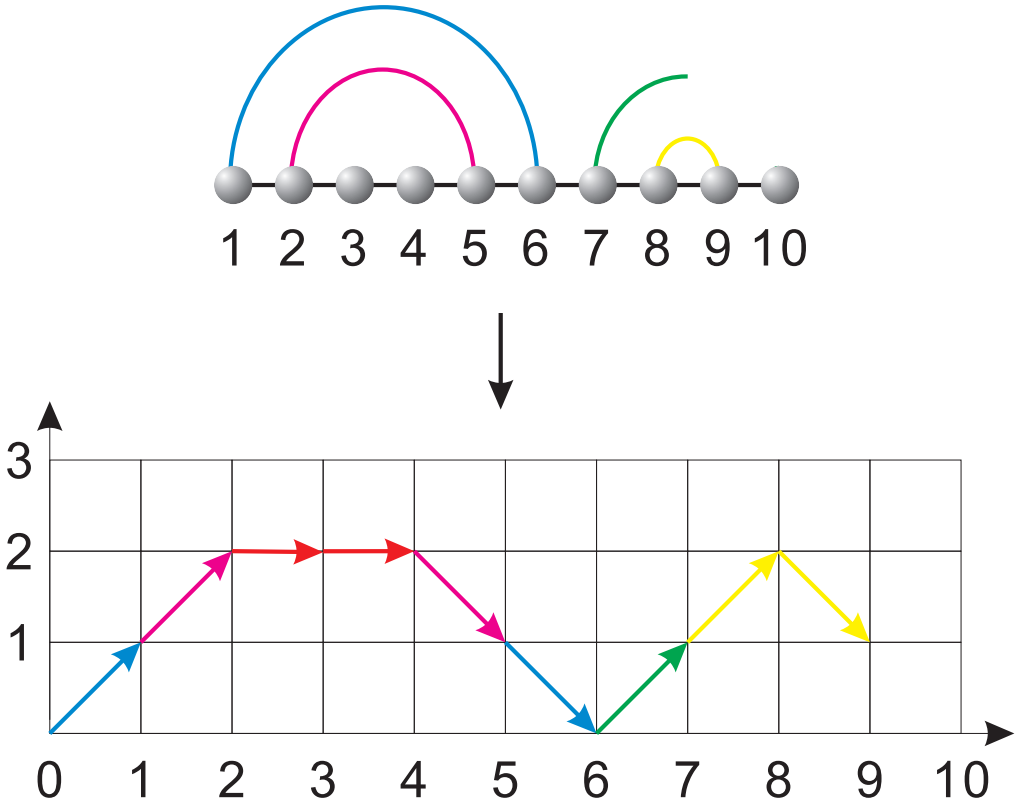


Figure: From a diagram to a path.

From a diagram to a path

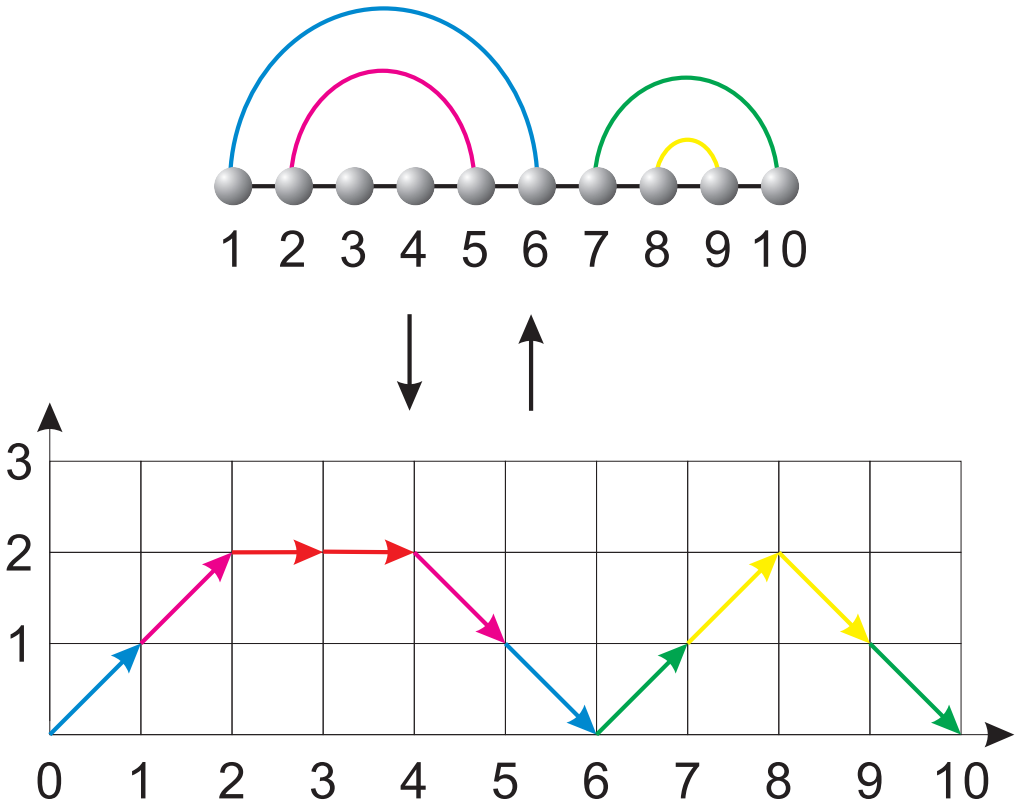


Figure: We obtain a path from $(0, 0)$ to $(10, 0)$.

Pseudoknot strucures

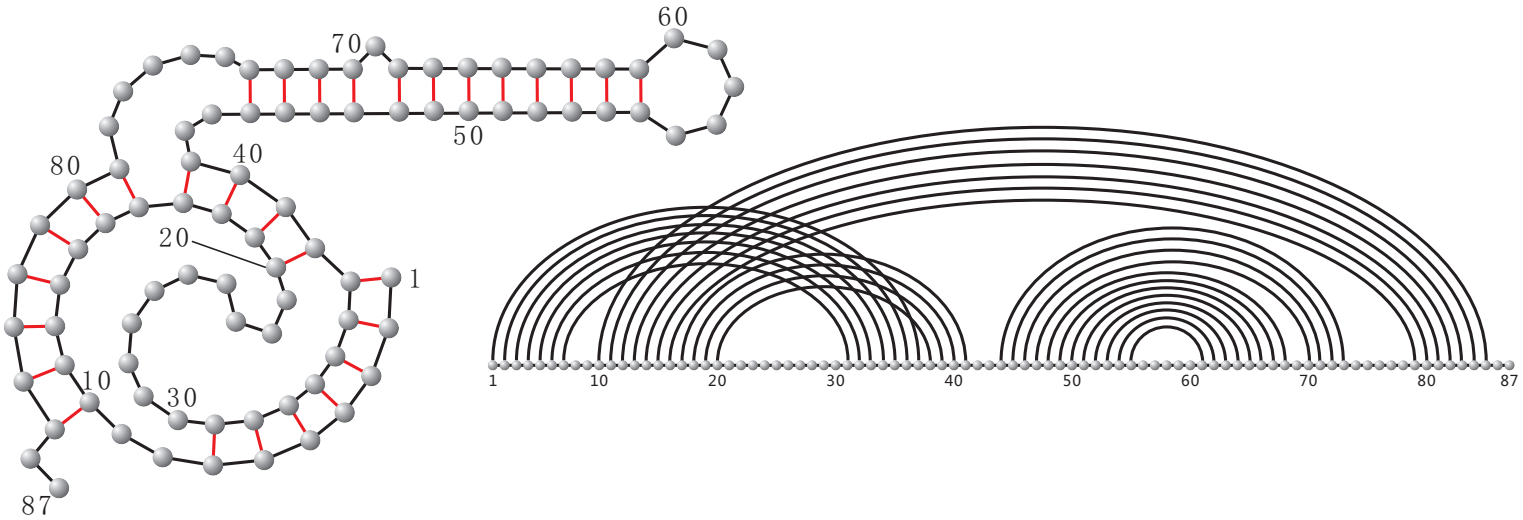
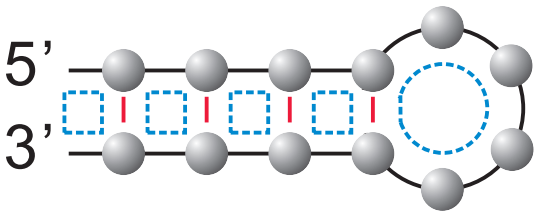


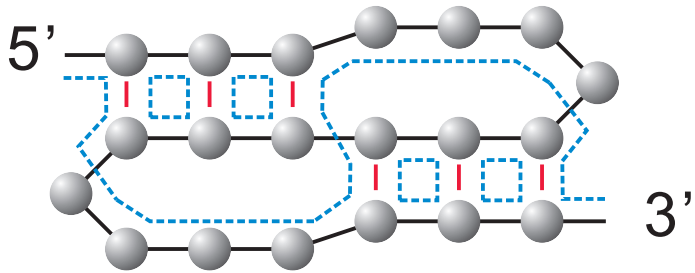
Figure: The Hepatitis Delta Virus (HDV)-pseudoknot structure represented as a planar graph and as a diagram: we display the structure as folded by the *ab initio* folding algorithm cross (left) and the diagram representation (right).

More than graphs...

- a secondary structure can be decomposed into “loops”,
- to specify a loop requires some kind of “orientation” i.e. how to turn around a vertex,
- its energy is loop-based depends on base pairs, bases and loop-type,
- pseudoknot structures have also a loop-decomposition
- energy is loop-based depends on base pairs, bases and loop-type or even simpler, when flat penalties of crossings are applied.



(A)



(B)

More than graphs: “fat”graphs

A **graph** consists of a set of half-edges, H, its **vertices** are **subsets** of half-edges and its **edges** are disjoint **pairs** of half-edges. A **fatgraph** consists of a set of half-edges, H its **vertices** are **cycles** of half-edges and its **edges** are disjoint **pairs** of half-edges. Thus, a fatgraph is given by (H, σ, α), where σ is the vertex-permutation and α a fixed-point free involution.

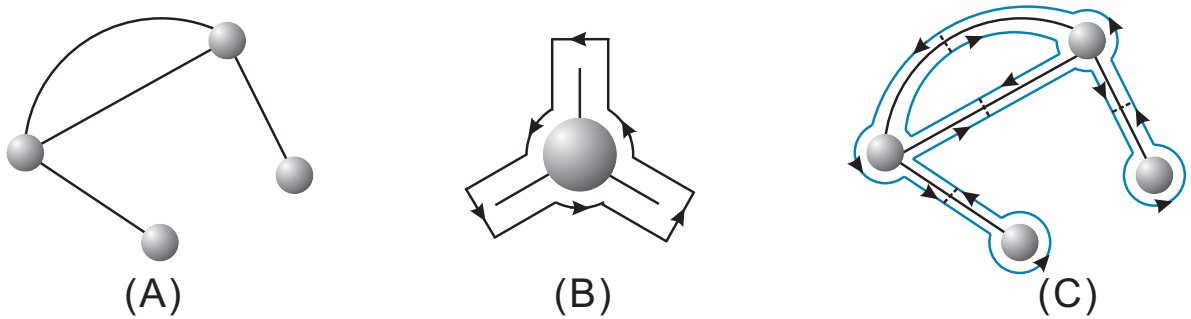


Figure: (A) a graph with 4 vertexes and 4 edges, (B) fattening of a vertex, (C) a fatgraph derived from (A). Any fatgraph induces a

Fatgraphs in the computer

A fatgraph having n edges can be presented by

- the vertex permutation σ and the involution α, representing the arcs,
- we can consider the permutation $\gamma = \alpha \circ \sigma$, whose cycles are called **boundary components**.

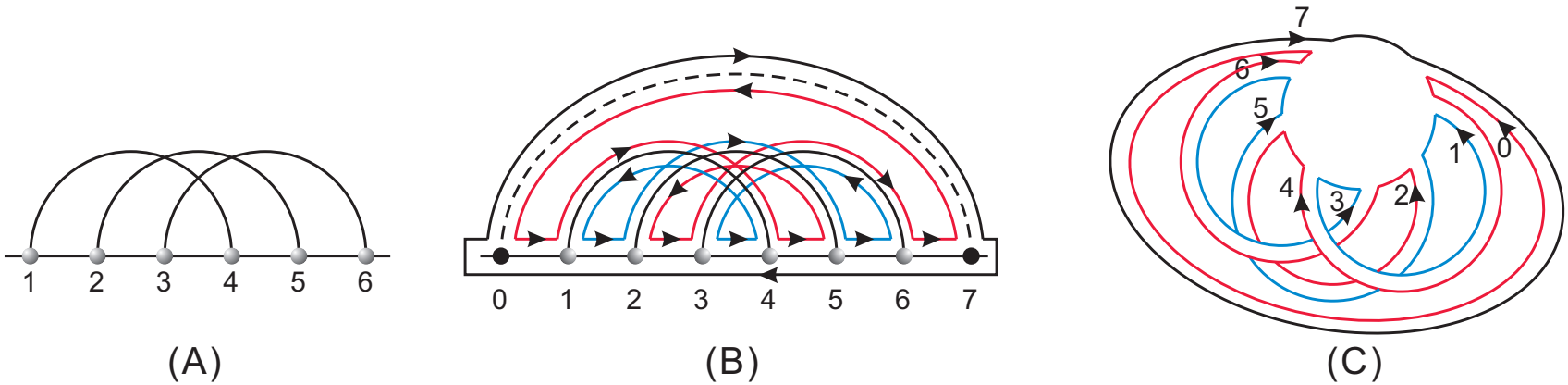


Figure: (A) A diagram, (B) a fatgraph of (A) augmented by an additional "rainbow" arc $(0, 7)$. (C) collapsing the backbone. Here $\gamma = \alpha \circ \sigma = (0, 4, 2, 6)(1, 5, 3)(7)$ has two cycles.

The Poincaré dual

Poincaré dual: Mapping a fatgraph (σ, α) to $(\alpha \circ \sigma, \alpha)$.

Reference: J Math Biol. 2012, Topological classification and enumeration of RNA structures by genus. Andersen JE, Penner RC, Reidys CM, Waterman MS.

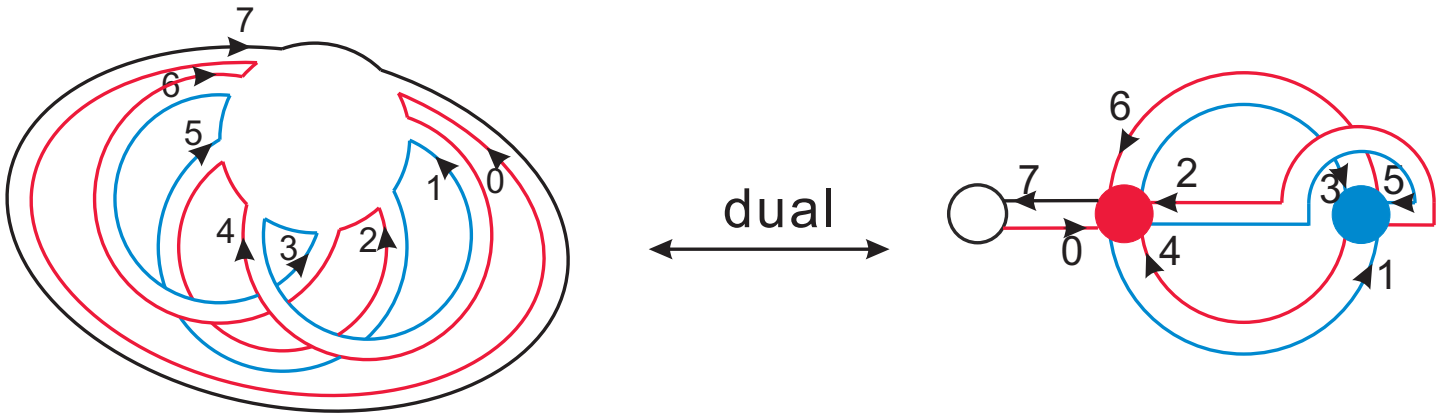


Figure: Bijection between a fatgraph with 1 vertex and 3 boundary components to a fatgraph with 3 vertexes and 1 boundary component.

Unicellular maps

A fatgraph with **one** boundary component is called a *unicellular map*. A *planted* unicellular map contains an additional vertex of degree one serving as its distinuished root.

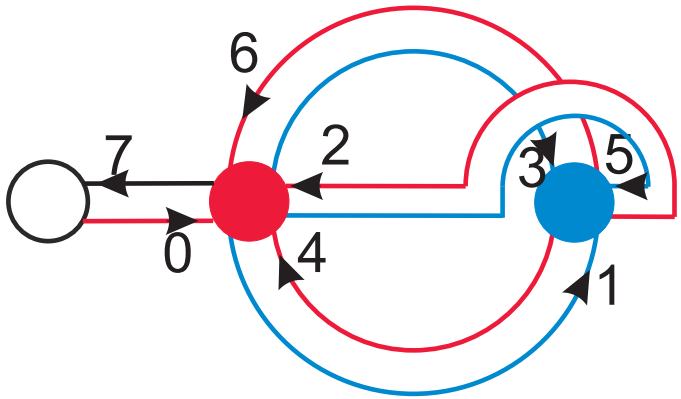


Figure: A unicellular map with three vertexes. The half-edges belonging to one vertex have the same color. Half-edges of a vertex appear in counterclockwise order, i.e., $(0, 4, 2, 6)$, $(1, 5, 3)$ and (7). The vertex (7) denotes the root of the rooted planar tree.

Two orders

- The tour of the unique boundary component. We write $a_1 <_\gamma a_2$ if a_1 appears before a_2 in this tour.
- The order of the half-edges induced by the vertex-cycle. We call $a_1 <_\sigma a_2$ if a_1 appears before a_2 counterclockwise in the vertex.

Figure: (A) The order of the half-edges appearing in the tour γ: $(0, 1, 2, 3, 4, 5, 6, 7)$. (B) The order of half-edges induces by the vertex-cycles: $(0, 4, 2, 6)$, $(1, 5, 3)$ and (7).

The genus

Let r denote the number of boundary components, v denote the number of vertices and e the number of edges. The genus of the fatgraph is given by Euler's characteristic formula

$$2 - 2g - r = v - e.$$

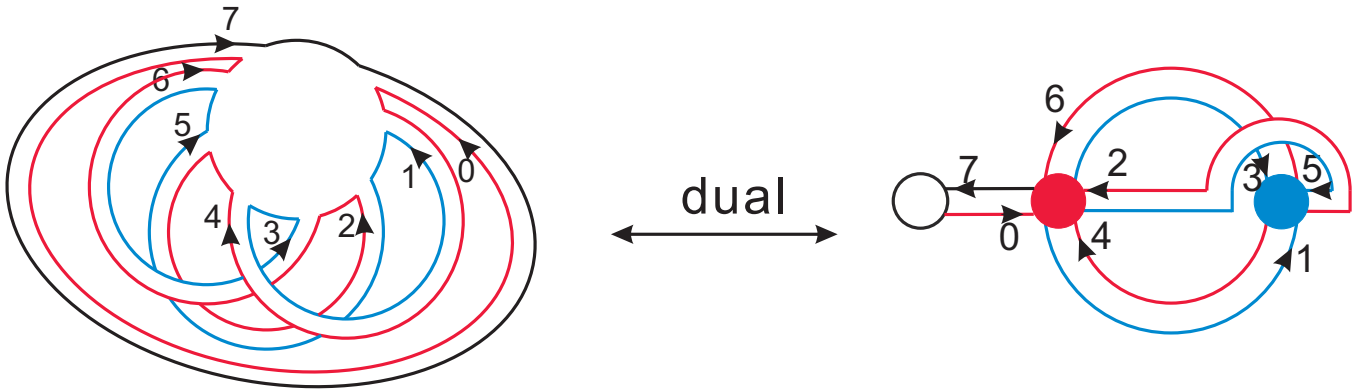


Figure: The Poincaré dual preserves the genus.

Gluing

- Consider three vertices v_1, v_2 and v_3, $v_i = (a_i, h_i^1, \ldots, h_i^{m_i})$, where a_i is the minimum labeled half-edge of v_i,
- set $\overline{v} = (a_1, h_2^1, \ldots, h_2^{m_2}, a_2, h_3^1, \ldots, h_3^{m_3}, a_3, h_1^1, \ldots, h_1^{m_1})$.

We consider $\overline{v}$ is obtained from by v_1, v_2 and v_3 by "gluing" as follows:

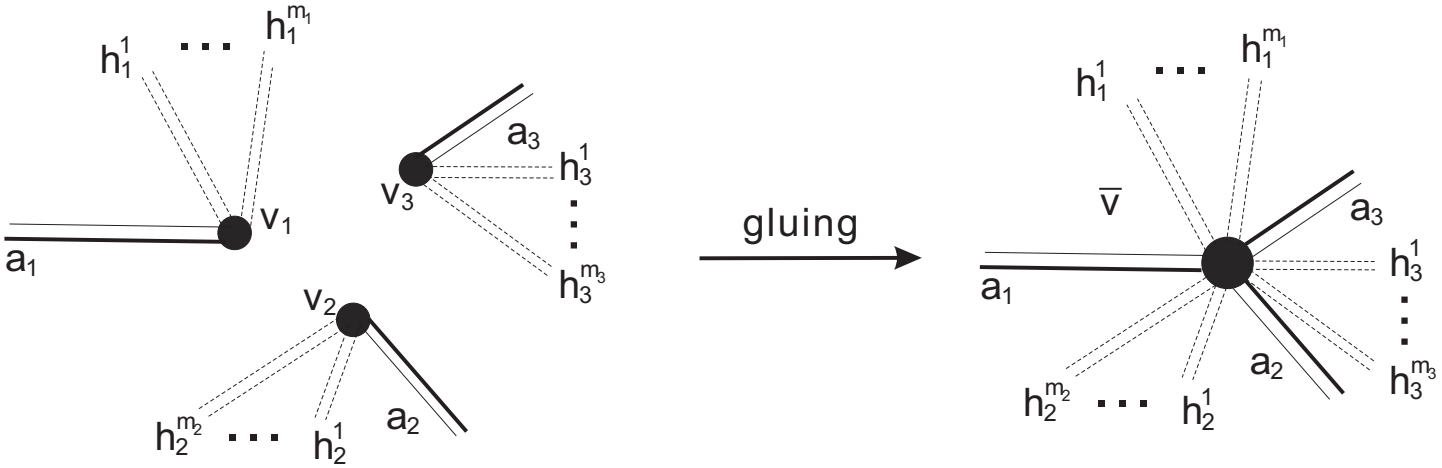


Figure: Gluing.

Slicing

- Consider $\overline{v} = (a_1, h_2^1, \ldots, h_2^{m_2}, a_2, h_3^1, \ldots, h_3^{m_3}, a_3, h_1^1, \ldots, h_1^{m_1})$,
- $\overline{v}$ is sliced into three vertices v_1, v_2 and v_3, where $v_i = (a_i, h_i^1, \ldots, h_i^{m_i})$.

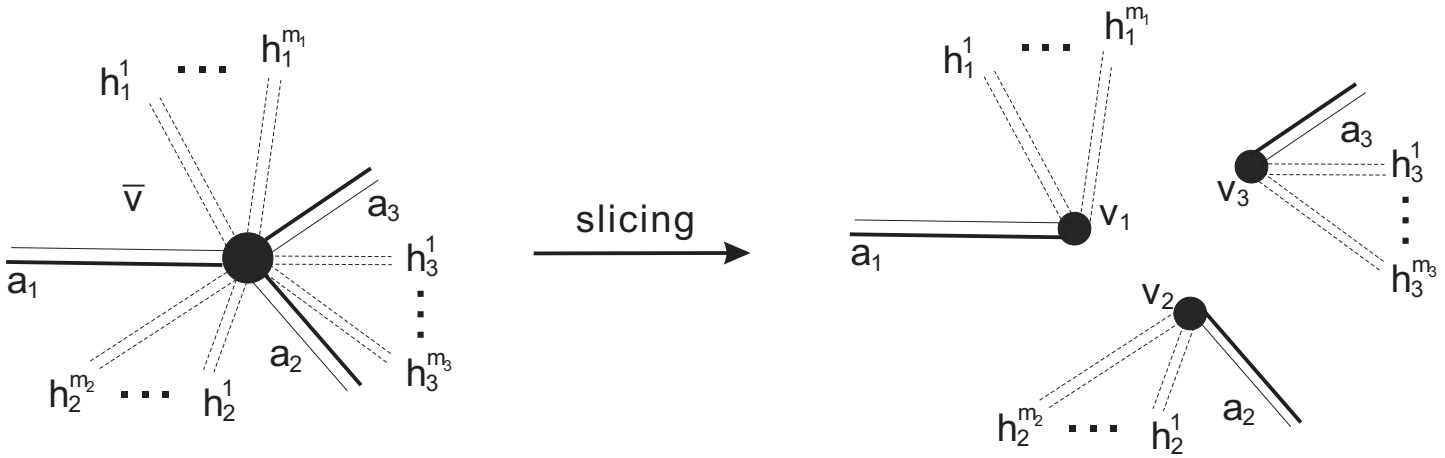


Figure: Slicing.

Intertwined

Definition

Three half-edges a_1, a_2 and a_3 are *intertwined* if

$$a_1 <_\sigma a_2 <_\sigma a_3, \quad a_1 <_\gamma a_3 <_\gamma a_2.$$

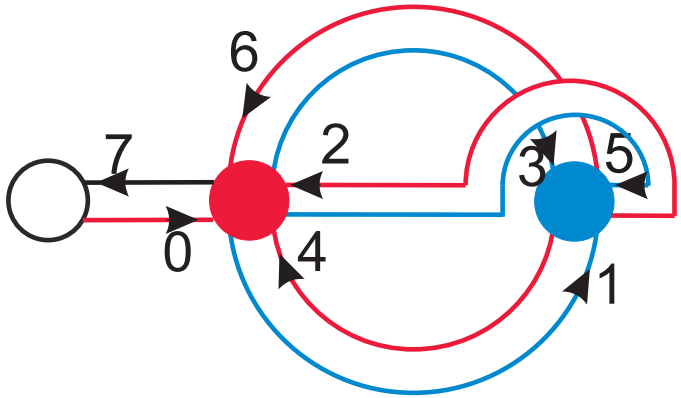


Figure: The half-edges 2, 4 and 6 in the vertex cycle $(0, 4, 2, 6)$ are intertwined. We have $2 <_\gamma 4 <_\gamma 6$ and $4 <_\sigma 2 <_\sigma 6$.

Gluing, slicing and intertwined

Lemma

Chapuy (2011) *Gluing three vertices of a fatgraph with one bdc of genus g generates a fatgraph with one bdc of genus $g+1$. Furthermore, it produces a vertex having three intertwined half-edges.*

Lemma

Chapuy (2011) *Slicing a vertex of a fatgraph with one bdc of genus $g+1$ with three intertwined half-edges generates a fatgraph with one bdc of genus g.*

Trisections

Lemma

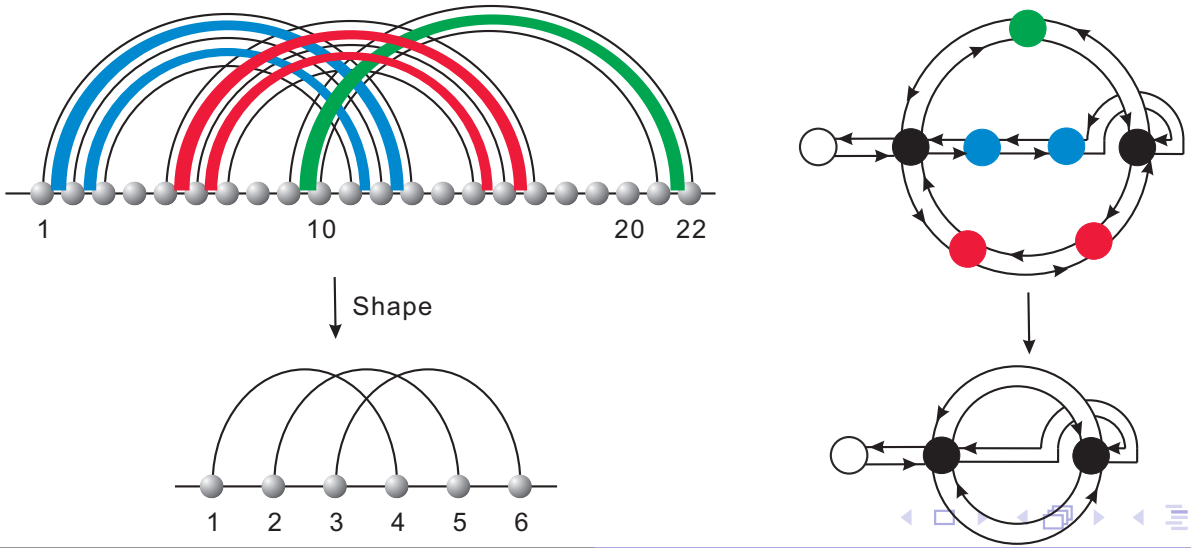
Chapuy (2011) *A planted unicellular map having n edges and genus g contains* $n+1$ *down-steps,* $n+1$ *up-steps and* $2g$ *trisections.*

Shapes

- remove any isolated vertices,
- replace each stack by a single arc.

Lemma

Reidys (2010) *There exist only finitely many shapes of fixed genus g.*



Theorem

(Huang-Reidys 2013) *The shape* **polynomial** *is given by*

$$\mathbf{S}_g(z) = \sum_{t=0}^{g-1} \kappa_t^{(g)} z^{2g+t}(1+z)^{2g+t}, \tag{1}$$

$$\sum_{\substack{0=g_0<g_1<\cdots<g_r=g \\ 0=t_0=t_1\leq t_2\leq\cdots\leq t_r=r-t}} \prod_{i=1}^{r} \frac{1}{2g_i} \binom{2g+t-(2g_{i-1}+(i-1))+t_i}{2(g_i-g_{i-1})+1}$$
$$\times Cat(2g+t-1)$$

and where $Cat(n) = \frac{1}{n+1}\binom{2n}{n}$ *and* $(t_i)_{r-t}$ *is a sequence of integers satisfying* $t_1 = 1$, $t_r = r-t$ *and* $t_i - t_{i-1} = 0$ *or* 1, $\forall 1 \leq i \leq r$.

Let $\epsilon_g(n)$ denote the number of unicellular map of genus g having n edges.

Corollary

$$2g \cdot \epsilon_g(n) = \binom{n+1-2(g-1)}{3}\epsilon_{g-1}(n) + \cdots + \binom{n+1}{2g+1}\epsilon_0(n). \quad (2)$$

Here the $2g$-factor on left hand side counts the number of trisection in $\mathfrak{m}_g$ and the binomial coefficients on the right counts the number of distinct selections of subsets of $(2k+1)$ vertices from $\mathfrak{m}_{g-k}$.

$$\epsilon_g(n) = \sum_{0=g_0<g_1<\cdots<g_r=g} \prod_{j=1}^{r} \frac{1}{2g_i}\binom{n+1-2g_{i-1}}{2(g_i-g_{i-1})+1} \cdot \epsilon_0(n), \quad (3)$$

Slice/glue paths

Definition

Suppose $\mathfrak{m}_g$ is a unicellular map of genus g having n edges. Then a sequence unicellular maps

$$(\mathfrak{m}^0 = \mathfrak{m}_{g_0=0}, \mathfrak{m}^1 = \mathfrak{m}_{g_1}, \ldots, \mathfrak{m}^r = \mathfrak{m}_{g_r=g})$$

is called a slice path from $\mathfrak{m}_g$ to $\mathfrak{m}_0$ and a glue path when considered from $\mathfrak{m}_0$ to $\mathfrak{m}_g$, where $\Xi(\mathfrak{m}_{g_i}, \tau_i) = (\mathfrak{m}_{g_{i-1}}, V_{g_{i-1}})$ holds for some τ_i in $\mathfrak{m}_{g_i}$, $0 < i \leq r$.

Counting glue-paths

We next consider $P_g(\mathfrak{m}^0)$, the set of distinct glue paths from a given $\mathfrak{m}^0 = \mathfrak{m}_0$ to some unicellular maps of fixed genus g.

Lemma

The cardinality of $P_g(\mathfrak{m}^0)$ is

$$\sum_{\substack{0=g_0<g_1<\cdots<g_r=g \\ \text{for some } r}} \prod_{i=1}^{r} \frac{1}{2g_i} \binom{n+1-2g_{i-1}}{2(g_i-g_{i-1})+1}.$$

Uniform generation

Lemma

(Huang-Reidys 2013) *There exists a linear time algorithm that generates a glue path p_g with probability $\epsilon_0(n)/\epsilon_g(n)$. Since a tree having n edges can be uniformly generated with probability $1/\epsilon_0(n)$, a matching of genus g having n edges can be generated uniformly.*

Reference: Generation of RNA pseudoknot structures with topological genus filtration, MBS, 2013, F.W.D. Huang, M.E. Nebel, C.M. Reidys
http://authors.elsevier.com/sd/article/S0025556413001788

We generate a unicellular map of genus g having n edges with uniform probability splits into two parts:

- we first generate a planar tree $\mathfrak{m}_0$ with n edges with uniform probability,
- second we generate a glue path from $P_g(\mathfrak{m}^0)$ with uniform probability.

It is well-known how to implement the first step by a linear time sampler and it thus remains to present an linear time algorithm for the second step.

Uniform generation of interaction structures of fixed topological genus

Corollary

(Han-Reidys 2013)

$$
\begin{aligned}
&2(g+1)\mathrm{B}_g(n-1)\\
&=\sum_{1\leq i\leq g+1}\binom{n+1-2(g+1-i)}{2i+1}\mathrm{B}_{g-i}(n-1)\\
&+\sum_{0\leq g_1\leq g}\sum_{1\leq i\leq g_1}\left(\sum_{k\geq 1}^{2i}\binom{m+1-2(g_1-k)}{k}\times\right.\\
&\left.\binom{n-m-2(g+1-g_1-(i-k))}{2i+1-k}\right)\mathrm{D}_{g+1-i}(n-1)
\end{aligned}
\tag{4}
$$

Warming up: the identical permutation

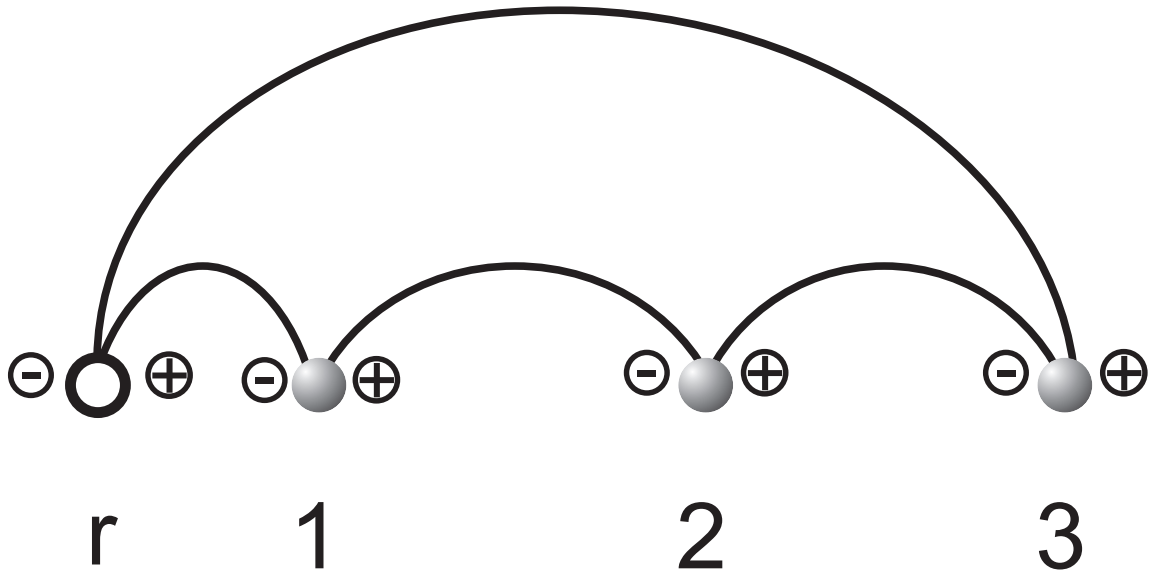


Figure: The identity permutation and its tangled diagram.

Constructing the tangled diagram

- given a signed permutation, a number is represented by a vertex,
- each such vertex has two half-edges. If the number it presents is positive, the labeling reads (left to right) $(-,+)$, and $(+,-)$, otherwise,
- there are arcs connecting vertices $(i,+)$ and $(i+1,-)$,
- arcs are untwisted if they connect two vertices of the same orientation and twisted, otherwise,

From tangled diagrams to *p*-diagram

- Given a tangle, first and the last vertex of its *p*-diagram are obtained from splitting the root.
- Each other *p*-diagram vertex is derived by splitting those in the tangled diagram each of them carries one half-edge.
- relabel the vertices in increasing order and connect two subsequent ones.

Lemma

There is a bijection betwee a tangled diagram of signed permutations and p-diagrams. In particular, there is a unique tangle for each p-diagram.

From tangles to *p*-diagrams

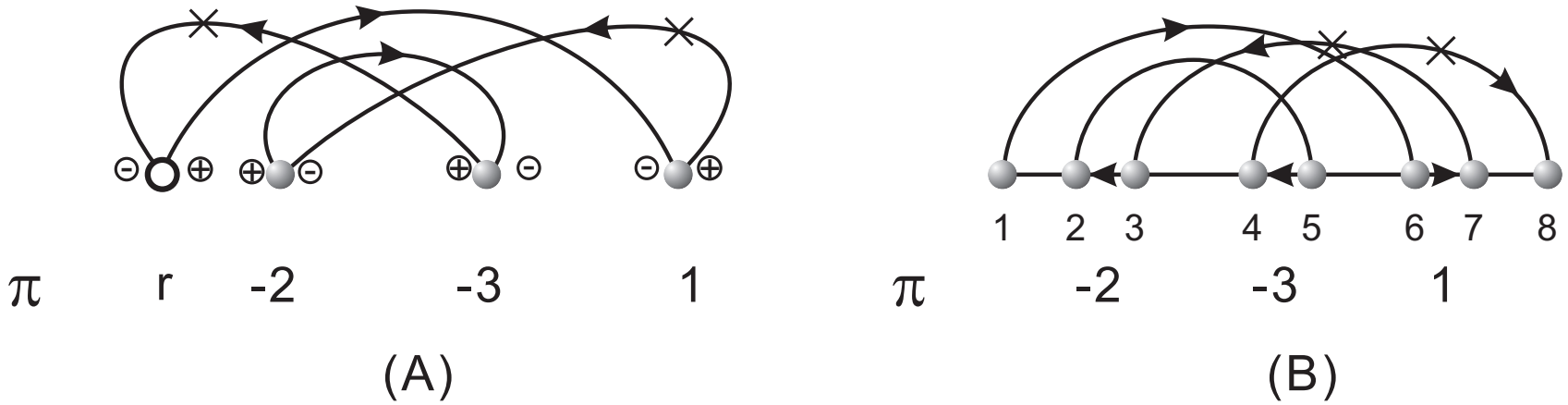


Figure: (A) A signed permutation $\pi = (-2, -3, 1)$ and its tangled diagram. The arc connected the root and 3 and the arc connected 1 and 2 are twisted (B) The p-diagram is induced by (A).

Fatgraphs

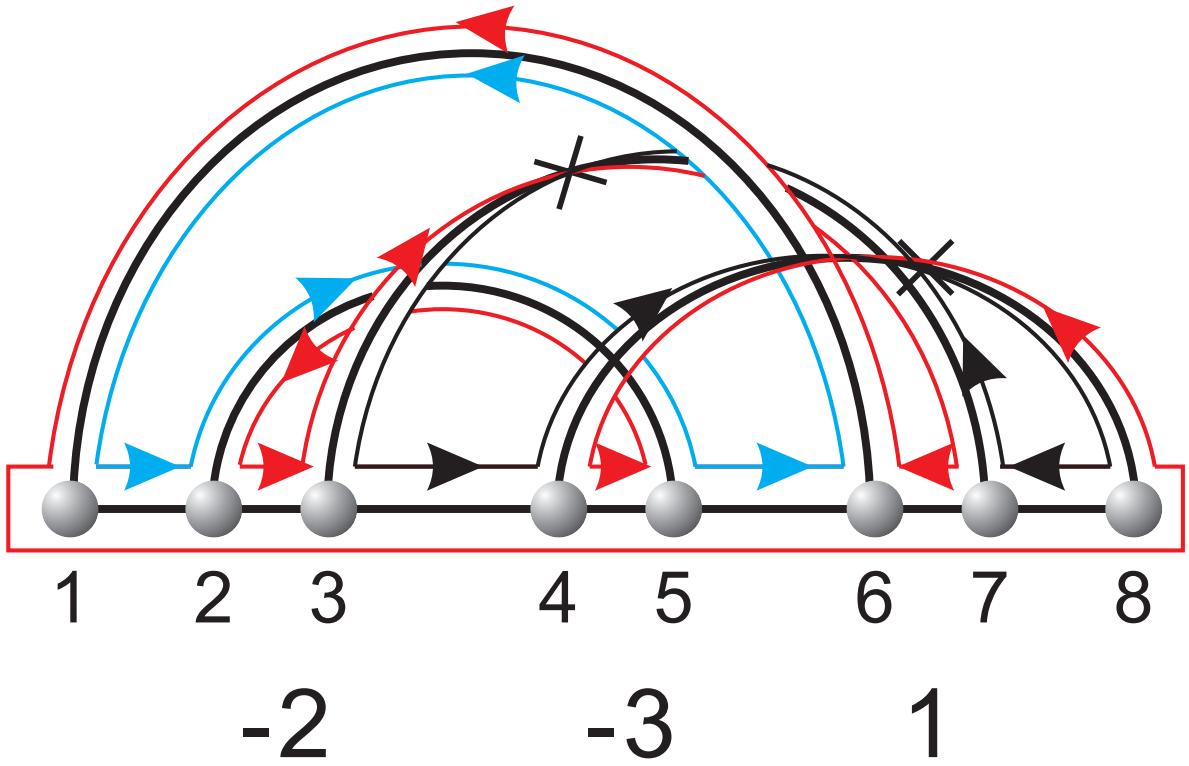


Figure: The p-diagram D_π of $\pi = (-2, -3, 1)$ and its fatgraph. Here the fatgraph consist of three boundary components (red, blue and black). The arrows on the boundaries indicates how to travel a

Topological genus

Euler characteristic and genus are given by

$$\chi(X_D) = v - e + r \tag{5}$$

$$g(X_D) = \begin{cases} 1 - \frac{1}{2}\chi(X_D), & \text{if } X_D \text{ is orientable,} \\ 1 - \chi(X_D) & \text{if } X_D \text{ is non-orientable.} \end{cases} \tag{6}$$

Setting $g(\pi) = 2g(X_D)$ if X_D is orientable *and*$g(\pi) = g(X_D)$ if X_D is non-orientable, $g(\pi)$ can be calculated by

$$g(\pi) = 1 - \frac{v - e + r}{2} = \frac{(n+1) - r + 1}{2}, \tag{7}$$

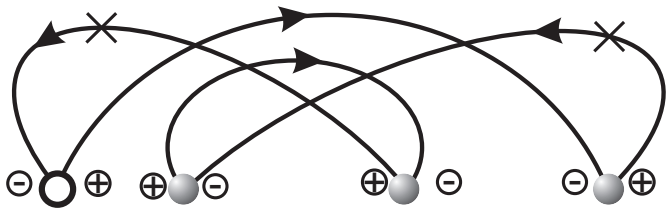
where n is the length of π and r is the number of boundary components.

The canonical boundary component

Lemma

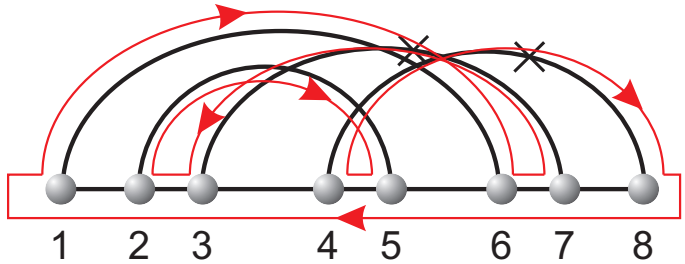
Let π denote a signed permutation and D is its p-diagram. Then the fatgraph X_D has has a unique boundary component, O^*, which travels all intervals $[2k, 2k+1]$, $\forall 1 \leq k \leq n$ in the p-diagram.

O^*, is called the *canonical boundary component*.



π r -2 -3 1

(A)



π -2 -3 1

(B)

Projection

Definition

Let π be a signed permutation and D_π is its p-diagram. Consider the three adjacent intervals $[2k-1, 2k]$, $[2k, 2k+1]$ and $[2k+1, 2k+2]$, where the middle interval belongs to O^*. If its adjacent intervals belong to the different boundary components, we call π_k in π **removable**. If π_k is removable we delete it from the tangle and relabel accordingly.

Lemma

A projection preserves the topological genus of a signed permutation.

Projection

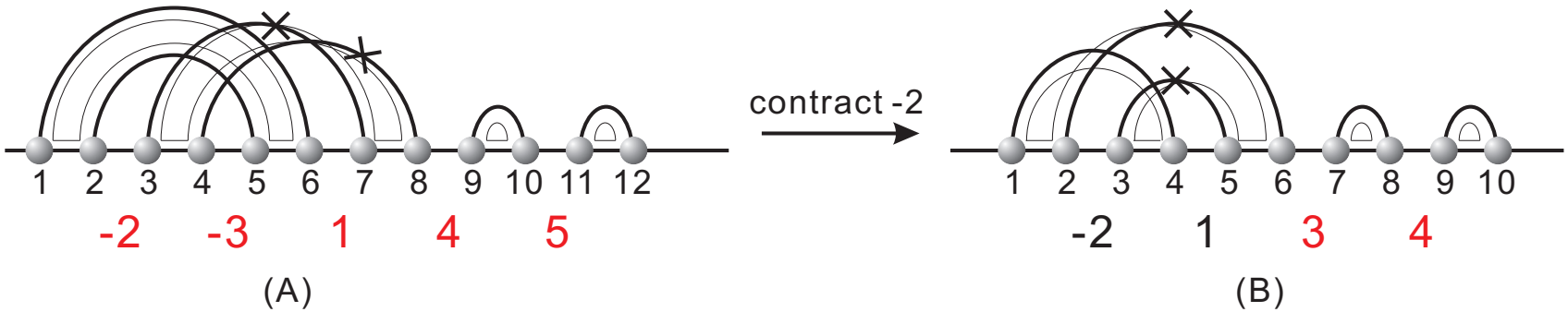


Figure: (A) a signed permutation $\pi = (-2, -3, 1, 4, 5)$ and its p-diagram. The p-diagram contains 6 arcs and 5 boundary components (O^* is not shown) so it has genus has genus $6 - 5 + 1 = 2$. All π_i are removable. (B) Projecting -2 from π we obtain $\pi' = (-2, 1, 3, 4)$ (relabeled) with genus $5 - 4 + 1 = 2$. It preserves the genus. However, $\pi_1 = -2$ and $\pi_2 = 1$ become not removable in the new permutation π.

Shapes

Definition

A signed permutation π is called a *p*-shape if it is fully contracted.

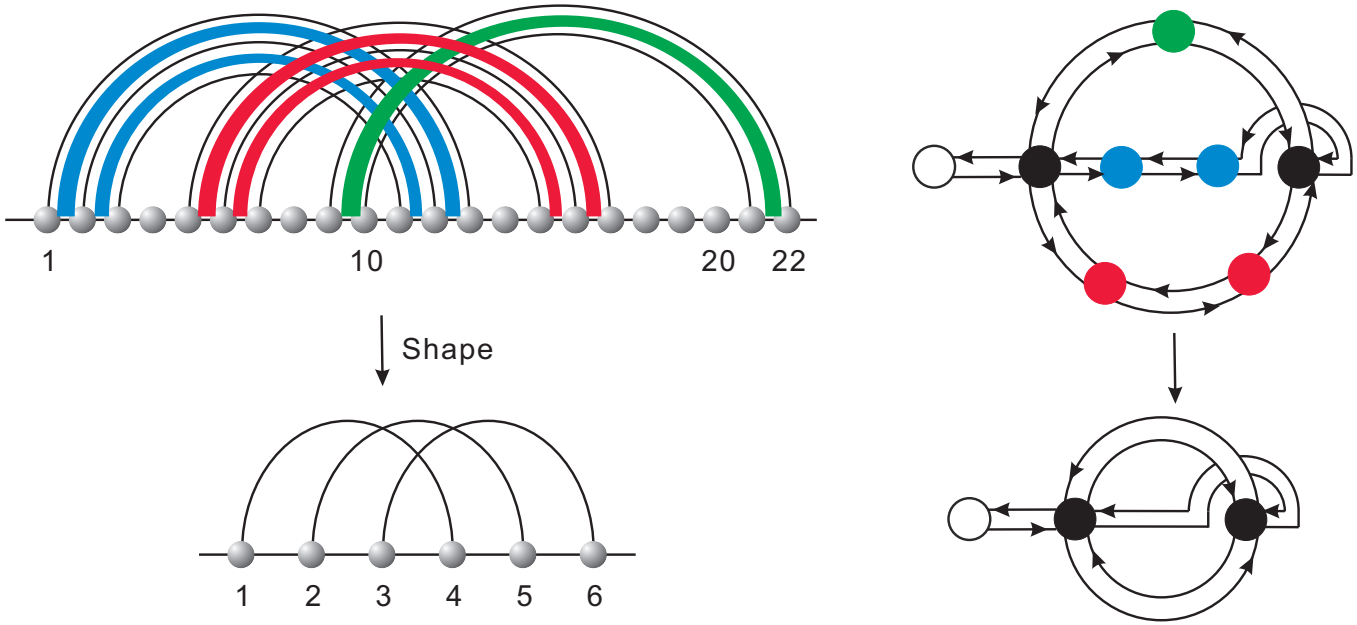


Figure: (A) A signed permutation $\pi = (-2, -3, 1, 4, 5)$ and its

Property of p-shapes

Lemma

If π is a p-shape then D_π contains only two boundary components and exactly n arcs. One is the canonical boundary component and the other travels all intervals $[2k-1, 2k]$, for all $1 \leq k \leq n$.

Lemma

Assume that π is a shape. If there is at least one twisted arc in D_π then the reversal distance of π is $d(\pi) = g(\pi)$ and $d(\pi) = g(\pi) + 1$, otherwise.

Reversal distance

Theorem (Hannenhalli, H. and Pevzner, P.)

The reversal distance of π is given by

$$d(\pi) = \begin{cases} b(\pi) - c(\pi) + h(\pi), & \text{if } \pi \text{ is not a fortress}, \\ b(\pi) - c(\pi) + h(\pi) + 1, & \text{if } \pi \text{ is a fortress}. \end{cases} \tag{8}$$

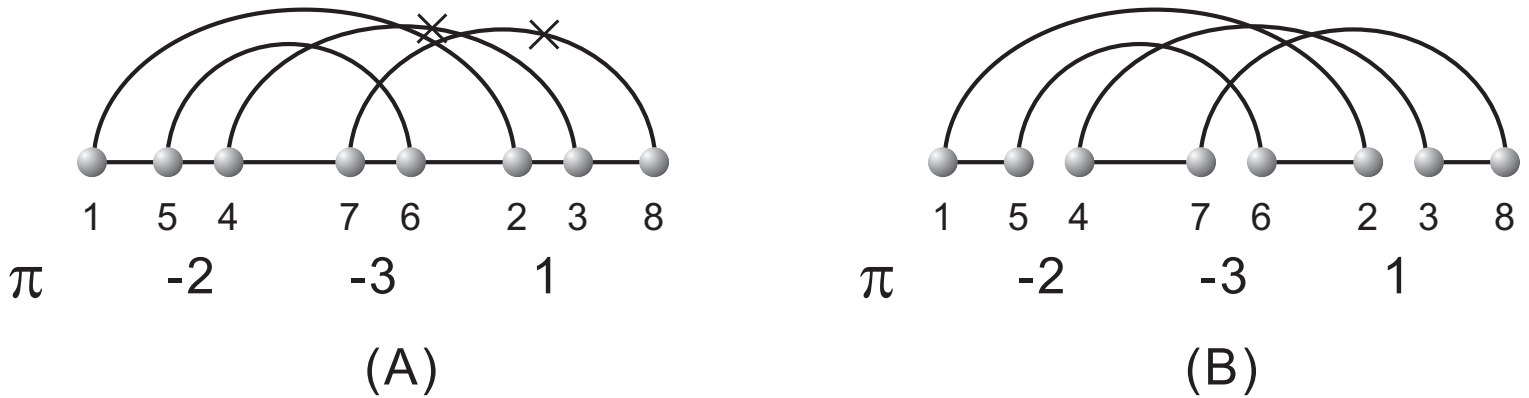


Figure: (A) The p-diagram of permutation $\pi = (-2, -3, 1)$ and (B) The break-point graph for the same permutation in (A). Here $b(\pi)$ is

The p-shape polynomial

Lemma

For a fixed $g > 0$, there are finitely many p-shapes.

Let $S(z)$ denote the generating function of p-shapes and s_g denote the number of p-shapes of genus g. Then we have

$$S(z) = \sum_{g>0} s_g z^g. \tag{9}$$

Distribution

Length	$g = 0$	1	2	3	4	5	6
0	1						
1	1	1					
2	1	3	4				
3	1	6	21	20			
4	1	10	65	160	148		
5	1	15	155	701	1620	1348	
6	1	21	315	2247	9324	19068	15104

Table: Distribution of signed permutations of fixed length by their associated topological genera.

Synopsis I

- Combinatorial structures contain often “more” data (e.g. loops in RNA). These can be captured by fattening the original graph, i.e. identifying a cell-complex whose graph is its defomation retract.
- A new paradigm namely “genus induction” naturally arises which allows to recursively construct more and more complex structures.
- The above concepts imply linear time uniform samplers for RNA structures and RNA interaction structures (not shown but similar)

Synopsis II

- Topological genus provides a new filtration of RNA structures and has led to topological folding algorithms (see Bioinformatics. 2011 Apr 15;27(8):1076-85)
- Analogue ideas apply for DNA rearrangements, where naturally non-orientable surfaces arise. For instance, Pevzner's reversal agorithm can be given a topological interpretation in terms of shapes.
- Genus induction suggests a normal form for any type of network.