1 Sortieren mit orientierten Reversionen

Wie groß ist die evolutionäre Distanz zwischen zwei gegebenen Spezies? Hierbei handelt es sich um eine zentrale Fragestellung der Biologie. Die evolutionäre Distanz ist im Allgemeinen unbekannt, jedoch gibt es auf biologischen Daten basierende mathematisch definierte Distanzmaße, um die tatsächliche Distanz zumindest anzunähern. Auf Genebene wird häufig die Edit-Distanz zwischen homologen Genen der zwei Spezies als Maß herangezogen, d.h.: Wie viele Mutationen werden benötigt, um die eine Gensequenz in die andere überzuführen?

Eine weitere Datenquelle findet sich auf Genomebene: Die Anordnung der Gene. Hier stellen wir uns die Frage, wie viele Genom- bzw. Chromosomenmutationen benötigt werden, um die Anordnung der Gene in Spezies A in die korrespondierende Anordnung der jeweils homologen Gene in Spezies B überzuführen. Im Folgenden wollen wir uns insbesondere dem Spezialfall widmen, dass die Genome der Spezies jeweils aus einem einzelnen linearen Chromosom bestehen. Einige andere Fragestellungen wie beispielsweise die, bei der die Spezies jeweils ein einzelnes zirkuläres Chromosom besitzen, sind zu dieser Fragestellung äquivalent. Weiterhin betrachten wir als zulässige Operation ausschließlich Inversionen (bzw. Reversionen) von Chromosomenabschnitten. Wir wollen nun eine mathematische Formulierung dieser Problemstellung anstreben.

1.1 Permutationen

Definition. *Eine orientierte Permutation mit n Elementen ist ein n-Tupel π mit* $\{|\pi_i| \mid i \in [n]\} = [n]$.

Das heißt, in einer orientierten Permutation der Größe n kommt jedes Element aus $[n]$ entweder mit positivem oder mit negativem Vorzeichen genau einmal vor. Wenn im Folgenden von eine Permutation die Rede ist, sei darunter stets eine orientierte Permutation zu verstehen.

Indem wir die Permutationen mit n Komponenten als Abbildungen von $[n] \cup (-[n])$ in sich selbst auffassen und als Verknüpfung die Hintereinanderausführung definieren, erhalten wir mit $\mathbf{id}_n := (1, \ldots, n)$ als neutralem Element die orientierte symmetrische Gruppe über n Elementen, in Zeichen $\bar{\mathbf{S}}_n$.

Definition. *Eine Abbildung $r_{i,j} : \bar{\mathbf{S}}_n \to \bar{\mathbf{S}}_n$ mit $1 \leq i \leq j \leq n$ nennen wir orientierte Reversion, wenn für alle $\pi \in \bar{\mathbf{S}}_n$ gilt:*

- *Für alle $k \in [n]$ mit $k < i$ oder $j < k$ gilt $r_{i,j}(\pi)_k = \pi_k$.*
- *Für alle $k \in [n]$ mit $i \leq k \leq j$ gilt $r_{i,j}(\pi)_k = -\pi_{j-k+i}$.*

Die orientierte Reversion $r_{i,j}$ dreht also innerhalb des Intervalls $[i, j]$ die Anordnung aller Elemente sowie deren Vorzeichen um.

So gilt beispielsweise: $r_{2,4}((1, -3, 2, 4, 5)) = (1, -4, -2, 3, 5)$.

Analog zu Permutationen werden wir im Folgenden unter einer Reversion stets eine orienterte Reversion verstehen.

Definition (Sorting by oriented Reversals)**.** *Das Problem „Sortieren von Permutationen mit orientierten Reversionen“ kann wie folgt formuliert werden:*

Gegeben: *Eine Permutation* $\pi \in \bar{\mathbf{S}}_n$.
Gesucht: *Eine kürzeste Folge von Reversionen* $(\varphi_1, \ldots, \varphi_k)$*, die* π *in die Permutation* $\mathbf{id}_n$ *überführt.*

Mit „Überführen“ ist natürlich $\varphi_k(\varphi_{k-1}(\ldots \varphi_1(\pi)\ldots)) = \mathbf{id}_n$ gemeint.

Wir halten folgende Beobachtung fest: Sei $\pi \in \bar{\mathbf{S}}_n$. Jede Operationsfolge, die die Permutation π in $\mathbf{id}_n$ überführt, führt auch $(\pi, n+1)$ in $\mathbf{id}_{n+1}$ über. Jede Operationsfolge, die $(\pi, n+1)$ in $\mathbf{id}_{n+1}$ überführt, induziert eine maximal so lange Operationsfolge, die π in $(\pi, n+1)$ überführt, indem stets das $n+1$-te Element ignoriert wird.

Im Folgenden betrachten wir eine eingeschränkte Variante des Problems, welche aber im Sinne dieser Beobachtung keine wirkliche Einschränkung darstellt:

Definition (Sorting by oriented Reversals (eingeschränkte Variante))**.** *Die eingeschränkte Variante von „Sortieren von zirkulären Permutationen mit orientierten Reversionen“ kann wie folgt formuliert werden:*

Gegeben: *Eine Permutation* $\pi \in \bar{\mathbf{S}}_n$ *mit* $\pi_n = n$.
Gesucht: *Eine kürzeste Folge von Reversionen* $(\varphi_1, \ldots, \varphi_k)$*, die* π *in die Permutation* $\mathbf{id}_n$ *überführt.*

1.2 Der Breakpoint-Graph

Um das Problem zu lösen, bedienen wir uns des Breakpointgraphen der Permutation. Die Struktur dieses Graphen verrät viel über seine zugrundeliegende Permutation.

Definition (Breakpointgraph)**.** *Der Breakpointgraph* $G(\pi) = (V, A)$ *der Permutation* $\pi \in \bar{\mathbf{S}}_n$ *ist ein ungerichteter kantengefärbter Multigraph mit folgenden Eigenschaften:*

$$
\begin{aligned}
V &= \{-|\pi_i|, +|\pi_i| \mid i \in [n]\} \\
A_O &:= \{\{-|\pi_i|, +|\pi_i|\} \mid i \in [n]\} \\
A_R &:= \{\{+|\pi_i|, -|\pi_{(i \bmod n)+1}|\} \mid i \in [n] \wedge \pi_i > 0 \wedge \pi_{(i \bmod n)+1} > 0\} \\
&\quad \cup \{\{-|\pi_i|, +|\pi_{(i \bmod n)+1}|\} \mid i \in [n] \wedge \pi_i < 0 \wedge \pi_{(i \bmod n)+1} < 0\} \\
&\quad \cup \{\{-|\pi_i|, -|\pi_{(i \bmod n)+1}|\} \mid i \in [n] \wedge \pi_i < 0 \wedge \pi_{(i \bmod n)+1} > 0\} \\
&\quad \cup \{\{+|\pi_i|, +|\pi_{(i \bmod n)+1}|\} \mid i \in [n] \wedge \pi_i > 0 \wedge \pi_{(i \bmod n)+1} < 0\} \\
A_D &:= \{\{+|\pi_i|, -((|\pi_i| \bmod n) + 1)\} \mid i \in [n]\} \\
A &= A_O \cup A_R \cup A_D
\end{aligned}
$$

Die Kanten aus der Menge A_O nennen wir Obverse-Edges. Diese repräsentieren die (gerichteten) Elemente der Permutation und werden meist rot (und gestrichelt) dargestellt. Die Kanten aus A_R sind die Reality-Edges. Sie repräsentieren die aktuellen Nachbarschaften der Permutationselemente und werden meist schwarz dargestellt. Die Kanten aus A_D werden als Desire-Edges bezeichnet. Sie stellen die angestrebten Nachbarschaften

dar, d.h. diejenigen Nachbarschaften, die die Elemente in sortiertem Zustand besitzen. Die Desire-Edges werden meist blau dargestellt. Jeder Knoten des Breakpoint-Graphen ist zu jeweils genau einer Obverse-, einer Reality- und einer Desire-Edge inzident.

Es ist zu beachten, dass der Breakpoint-Graph die Elemente der Permutation zyklisch auffasst. Das bedeutet, dass das Element π_n als direkter Vorgänger von π_1 interpretiert wird.

Ein Breakpoint-Graph ist in Abbildung 1 dargestellt.

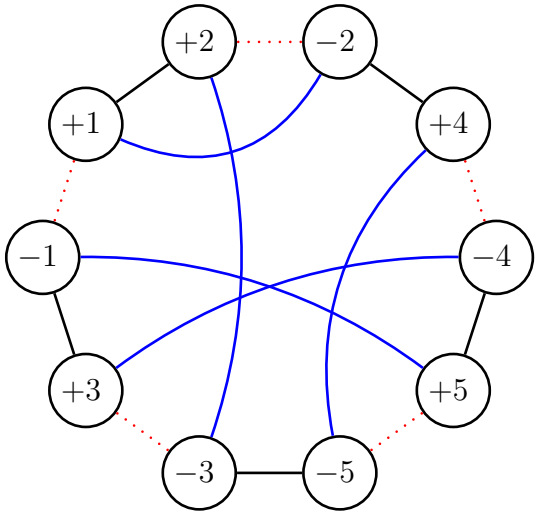

Abbildung 1: Dargestellt ist der Breakpoint-Graph der Permutation $(+4, +2, -1, -3, +5)$. Unter anderem die Permutationen $(-3, +5, +4, +2, -1)$ und $(-5, +3, +1, -2, -4)$ besitzen jedoch ebenfalls diesen Breakpoint-Graphen.

Den Breakpoint-Graphen einer Permutation π kann beispielsweise wie folgt erzeugt werden: Man geht die Permutation von vorne nach hinten durch und erzeugt für jedes Element π_i die beiden Knoten $+\pi_i$ und $-\pi_i$. Zwischen $+\pi_i$ und $-\pi_i$ kann sofort eine Obverse-Edge eingefügt werden. Wenn wir vom Element π_i zum Element π_{i+1} kommen, muss die passende Reality-Edge eingefügt werden. Hier kommt es darauf an, wie die jeweiligen Vorzeichen von π_i und π_{i+1} sind:

- Sind beide Elemente positiv, dann fügen wir eine Reality-Edge zwischen $+|\pi_i|$ und $-|\pi_{i+1}|$ ein.
- Sind beide Elemente negativ, dann fügen wir eine Reality-Edge zwischen $-|\pi_i|$ und $+|\pi_{i+1}|$ ein.
- Gilt $\pi_i > 0$ und $\pi_{i+1} < 0$, dann fügen wir eine Reality-Edge zwischen $+|\pi_i|$ und $+|\pi_{i+1}|$ ein.

- Gilt $\pi_i < 0$ und $\pi_{i+1} > 0$, dann fügen wir eine Reality-Edge zwischen $-|\pi_i|$ und $-|\pi_{i+1}|$ ein.

Dieses Prinzip wird natürlich auf geeignete Weise auch für π_n und π_1 angewendet. Die Desire-Edges können anschließend problemlos der Definition entsprechend eingefügt werden.

Lemma. *Für je zwei verschiedene Farben A und B aus {Reality, Desire, Obverse} zerfällt der Breakpoint-Graph einer Permutation in alternierende Kreise aus Kanten der Farben A und B.*

Definition.

- *Alternierende Kreise aus Reality- und Desire-Edges werden Reality-Desire-Kreise genannt.*
- *Alternierende Kreise aus Reality- und Obverse-Edges werden Reality-Obverse-Kreise genannt.*
- *Alternierende Kreise aus Desire- und Obverse-Edges werden Desire-Obverse-Kreise genannt.*

Beobachtung. *Der Breakpoint-Graph einer Permutation besitzt genau einen Reality-Obverse-Kreis und genau einen Desire-Obverse-Kreis.*

Anschaulich entspricht der Reality-Obverse-Kreis genau der aktuellen Abfolge der Elemente der Permutation, während der Desire-Obverse-Kreis der Abfolge der Elemente von $\mathbf{id}_n$ und $\mathrm{refl}(\mathbf{id}_n)$ bzw. deren zyklischen Verschiebungen entspricht. Eine zyklische Verschiebung einer Permutation entsteht dadurch, dass die Permutation in zwei Intervalle aufgespalten wird und diese Intervalle vertauscht werden. Eine zyklische Verschiebung von $\mathbf{id}_6$ hätte beispielsweise die Gestalt $(5, 6, 1, 2, 3, 4)$.

Der Breakpoint-Graph wird meist so dargestellt, dass die Knoten auf einem gedachten Kreis liegen, welcher durch den Reality-Obverse-Kreis approximiert wird. Die Desire-Edges befinden sich innerhalb des Kreises, wobei sie zur besseren Hervorhebung zur Mitte des Kreises geschwungen sind.

Lemma 1. *Sei $G(\pi)$ der Breakpoint-Graph einer Permutation $\pi \in \bar{\mathbf{S}}_n$.*

- *$G(\pi)$ enthält genau dann exakt n Reality-Desire-Kreise, wenn $\pi = \mathbf{id}_n$ oder $\pi = \mathit{refl}(\mathbf{id}_n) = (-n, \ldots, -1)$ oder wenn π eine zyklische Verschiebung von $\mathbf{id}_n$ oder $\mathit{refl}(\mathbf{id}_n)$ ist.*
- *Andernfalls enthält $G(\pi)$ weniger als n Reality-Desire-Kreise.*

Aus diesem Lemma geht hervor, dass es eine notwendige Bedingung darstellt, die Zahl der Reality-Desire-Kreise zu maximieren, wenn wir die Permutation sortieren wollen. In einem Breakpoint-Graph mit $2n$ Knoten kann es maximal n Reality-Desire-Kreise geben. Einen Breakpoint-Graphen mit n solcher Kreise bezeichnen wir als „sortiert“. In Abbildung 2 ist der sortierte Breakpoint-Graph der Permutation $\mathbf{id}_5$ dargestellt.

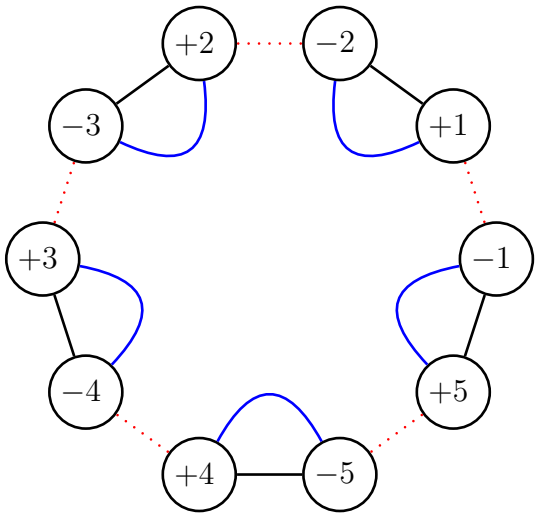

Abbildung 2: Dargestellt ist der Breakpoint-Graph der Permutation $\mathbf{id}_5 = (+1, +2, +3, +4, +5)$.

Definition. *Eine Reversion auf dem Breakpoint-Graphen einer Permutation ist eine Graphoperation, bei der zwei verschiedene Reality-Edges entfernt und zwei neue Reality-Edges hinzugefügt werden, sodass der resultierende Graph der Breakpoint-Graph einer Permutation ist.*

Zwei entfernte Reality-Edges legen hierbei die beiden neuen Reality-Edges eindeutig fest. Betrachte hierzu Abbildung 3.

Definition. *Eine Folge von Reversionen auf dem Breakpoint-Graphen nennen wir dann sortierend, wenn sie die Zahl der Reality-Desire-Kreise maximiert.*

Die zu einer Reversion ρ auf dem Breakpoint-Graphen korrespondierenden kanonische Reversionen auf der Permutation ist wie folgt definiert: Die Entfernung zweier Reality-Edges durch ρ teilt den Reality-Obverse-Kreis in zwei Reality-Obverse-Bögen, die jeweils mit einer Obverse-Edge beginnen und enden. Von diesen Bögen korrespondiert mindestens einer zu einem zusammenhängenden Intervall der Permutation.

Fall 1. Angenommen, einer der Bögen entspricht einem Intervall $(\pi_1, \ldots, \pi_j)$ mit $j < n$. Die zur Reversion auf dem Breakpoint-Graphen korrespondierende Reversion auf der Permutation ist dann $r_{1,j}$.

Fall 2. Angenommen, einer der Bögen entspricht einem Intervall $(\pi_i, \ldots, \pi_j)$ mit $1 < i$ und $j < n$. Die korrespondierende Reversion auf der Permutation ist dann $r_{i,j}$.

Fall 3. Angenommen, einer der Bögen entspricht einem Intervall $(\pi_i, \ldots, \pi_n)$ mit $1 < i$. Dann entspricht der zweite Bogen dem Intervall $(\pi_1, \ldots, \pi_{i-1})$. Dies ist aber gerade der erste Fall, sodass der dritte Fall ignoriert werden kann.

Jede Reversion auf dem Breakpoint-Graphen lässt sich damit einem der ersten beiden Fälle zuordnen. Die notwendige Bedingung $\rho(G(\pi)) = G(r_{i,j}(\pi))$ ist dabei stets erfüllt. Eine wichtige Beobachtung ist, dass das Element π_n nie in dem invertierten Intervall liegt (dies könnte nur dann auftreten, falls wir bei Fall 1 bzw. 3 das zweite Intervall statt dem ersten invertieren, was wir aber nicht tun). Falls $\pi_n = n$ gilt, korrespondiert jede Reversionsfolge auf dem Breakpoint-Graphen, die diesen sortiert, zu einer Reversionsfolge auf der Permutation, die diese tatsächlich in die Identität $\mathbf{id}_n$ überführt. Das ist der Fall, da $\mathbf{id}_n$ die einzige Permutation ist, die zum sortierten Breakpoint-Graphen korrespondiert und die gleichzeitig das Element n an der n-ten Position besitzt. Damit wird auch klar, weshalb wir die eingeschränkte Variante des Problems betrachten: Für jede Inputpermutationen π gilt hier $\pi_n = n$.

In Abbildung 3 ist eine Reversion auf einem Breakpoint-Graphen und die korrespondierende Reversion auf der Permutation dargestellt.

Um eine Reversion ρ auf einem Breakpoint-Graphen $G(\pi)$ in eine Reversion $r_{i,j}$ für π mit $\pi_n = n$ zu übersetzen gibt es mehrere Möglichkeiten. So kann man entweder eine ausgeklügelte Datenstruktur für den Breakpoint-Graphen verwenden, die ein Mapping zwischen Permutation und Breakpoint-Graph mit $O(\log n)$ Aufwand pro Reversion ermöglicht. Alternativ erzeugt man nach der Reversion auf $G(\pi)$ die eindeutige zum Breakpoint-Graphen $\rho(G(\pi))$ korrespondierende Permutation π' mit $\pi'_n = n$, für die $G(\pi') = \rho(G(\pi))$ gilt. Bei dieser Permutation handelt es sich dann um $\pi' = r_{i,j}(\pi)$. Ist man noch an den Indizes i, j interessiert, kann man diese leicht durch einen Vergleich von π und π' ermitteln.

Um aus einem Breakpoint-Graphen G die Permutation π zu erzeugen, für die $\pi_n = n$ und $G = G(\pi)$ gilt, geht man folgendermaßen vor: Zunächst ermittelt man den Knoten $+n$ und legt $\pi_n := n$ fest. Bei diesem Knoten beginnend läuft man nun den Reality-Obverse-Kreis ab, wobei man mit einer Reality-Edge beginnt. Immer wenn man eine Reality-Edge abgelaufen hat, wird ein neues Element der Permutation festgelegt. Wenn die i-te Reality-Edge abgelaufen wurde und man dabei den Knoten x besucht, definiert man $\pi_i := +|x|$, falls $x < 0$ und andernfalls $\pi_i := -|x|$. Wenn alle Elemente ermittelt wurden, ist man fertig.

Man beobachte, dass bis auf eine einzige Ausnahme jede Reversionen auf der Permutation in eine korrespondierende Reversion auf dem Breakpoint-Graphen übersetzt werden kann. Die Ausnahme ist die Reversion $r_{1,n}$. Der Hintergrund liegt darin, dass bei Permutationen die Leserichtung eine Rolle spielt (d.h. ob wir die Elemente von links nach rechts oder umgekehrt betrachten), jedoch eine Permutation π und deren Spiegelung $\text{refl}(\pi)$ stets denselben Breakpoint-Graph besitzen. Da - wie wir später sehen werden - die Zahl der Reversionen, die wir zur Sortierung der Permutation benötigen, allein von der topologischen Struktur des korrespondierenden Breakpoint-Graphen abhängt, wird ersichtlich, dass die Reversion $r_{1,n}$ nie Teil einer optimalen Operationsfolge sein kann.

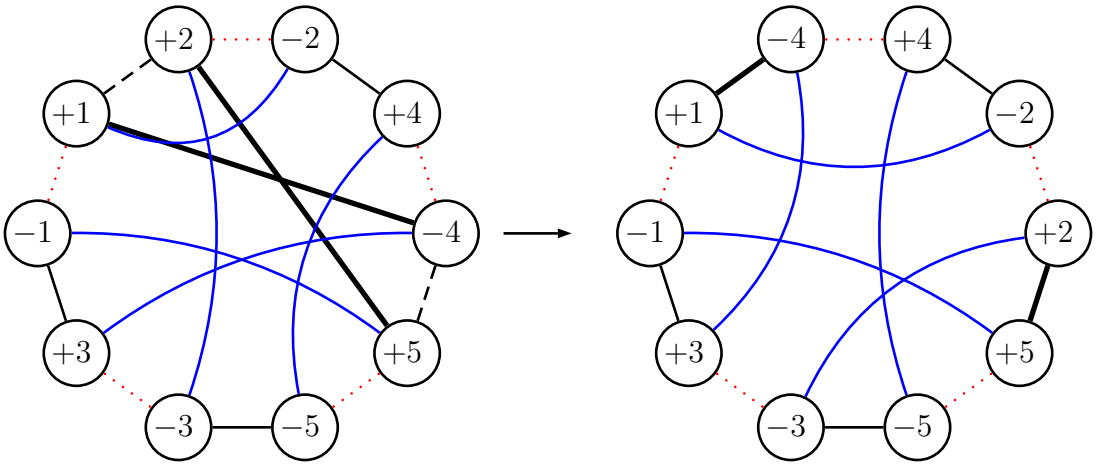

Abbildung 3: Links befindet sich der Breakpoint-Graph $G(\pi)$ der Permutation $\pi := (+4, +2, -1, -3, +5)$. Hierbei wird gerade eine Reversion angewandt, die die gestrichtelt dargestellten Reality-Edges $\{+1, +2\}$ und $\{+5, -4\}$ entfernt und die fett dargestellten Reality-Edges $\{+1, -4\}$ und $\{+2, +5\}$ einfügt. Man beachte, dass dies die einzige Möglichkeit ist, zwei neue Reality-Edges einzufügen. Die andere Möglichkeit resultiert in einem Breakpoint-Graphen, der zwei Reality-Obverse-Kreise besitzt und damit nicht zu einer Permutation gehört. Die zu der dargestellten Reversion auf $G(\pi)$ korrespondierende Reversion auf π ist $r_{1,2}$. Wird diese Reversion auf π angewendet, erhalten wir $r_{1,2}(\pi) = (-2, -4, -1, -3, +5)$. Auf der rechten Seite ist der Breakpoint-Graph $G(r_{1,2}(\pi))$ zu sehen, d.h. der Breakpoint-Graph nach der Reversion.