

Bioinformatik

Computational Molecular Biology

Prof. Dr. Volker Sperschneider

- Theoretische Informatik -

Fachbereich Mathematik/Informatik

Universität Osnabrück

(1) Vorbemerkungen

- Biologie als Leitwissenschaft des 21. Jahrhunderts und Informatik als methodische Fundamentalwissenschaft
- Biotechnologie, Genomprojekt, Daten(sint)flut
- Rechnen statt Messen (Data Mining)
- Komplexität von Problemen

(2) Drei ausgewählte Probleme der Bioinformatik und ihre mathematisch/informatische Lösung

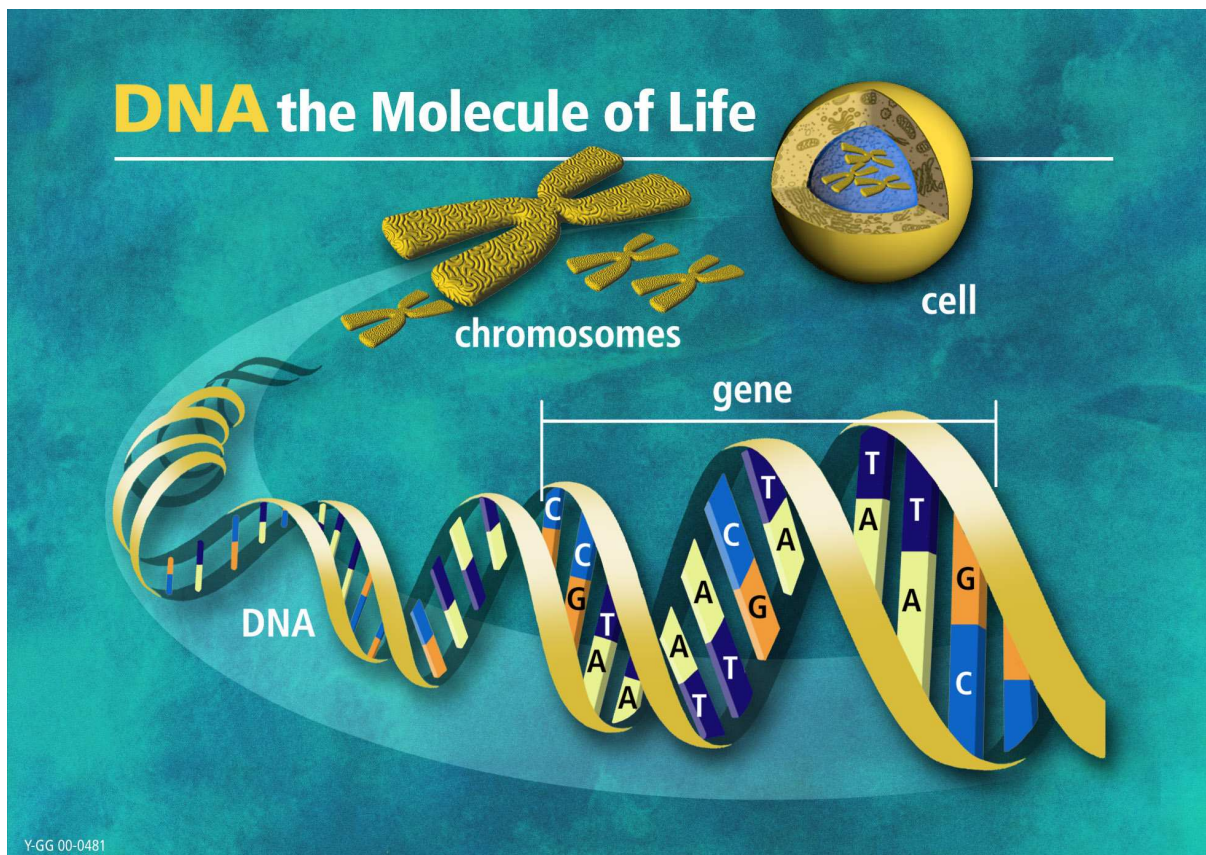
- Stammbaumrekonstruktion
- Genomtransformation
- Sequenzsuche

(1) Vorbemerkungen

DNA des Menschen:

In jeder Zelle eine **Doppelhelix** mit ca.
3.000.000.000 Zeichen A(denin), C(ytosin),
G(uanin), T(hymin)

AACGCGTTAATGATTGTGTTGAACACTGCGT...



Vergleich:

Bibliothek mit 1000 Büchern mit je 1000 Seiten

Riesige und zahlreiche DNA- und Protein-Datenbanken

Aus biologischen Anliegen erwachsen sehr schnell algorithmische Probleme. Beispiele auf verschiedenen Ebenen:

- Grobe Kartierung des Erbgutes: Mapping
- Exakte Bestimmung des Erbgutes: Sequencing
- Struktur des Erbgutes: Gene und Signale
- Genomvergleich und evolutionäre Distanz
- Stammbaumrekonstruktion

Algorithmische Probleme können

- gelegentlich ganz einfach und in naheliegender Weise klassisch gelöst werden („**analytische Formel**“)
- manchmal unter massivem Einsatz von Rechnern gelöst werden („**Suchalgorithmen**“)
- häufig aber nur mittels umfassender mathematischer Kompetenz und durch Aufbieten nichttrivialer algorithmischer Ideen gelöst werden („**no hacking**“)
- oft überhaupt nicht algorithmisch effizient gelöst werden, wobei genau dies auch noch bewiesen werden kann („**unlösbare Probleme**“)

Es ist gut zu wissen, ob ein Problem grundsätzlich unlösbar ist: Dann versucht man nicht in sinnloser Weise, nach nicht existenten Lösungen zu suchen.♣

♣ Vergleich: Quadratur des Kreises, Dreiteilung des Winkels, Perpetuum Mobile.

(2) Drei ausgewählte Probleme der Bioinformatik und ihre Lösung

Thema 1: Stammbaumrekonstruktion

Heute lebende (erfundene) Spezies seien

- Aliens
- Brainers
- Chunks
- Doorts
- Eymen
- Fitfats
- Gnomaden

Aufgrund von Gen-Vergleichen (behandeln wir im nächsten Abschnitt) sei bekannt, welche evolutionären Zeiten verstrichen sind, seit sich je zwei Spezies auseinander entwickelt haben:

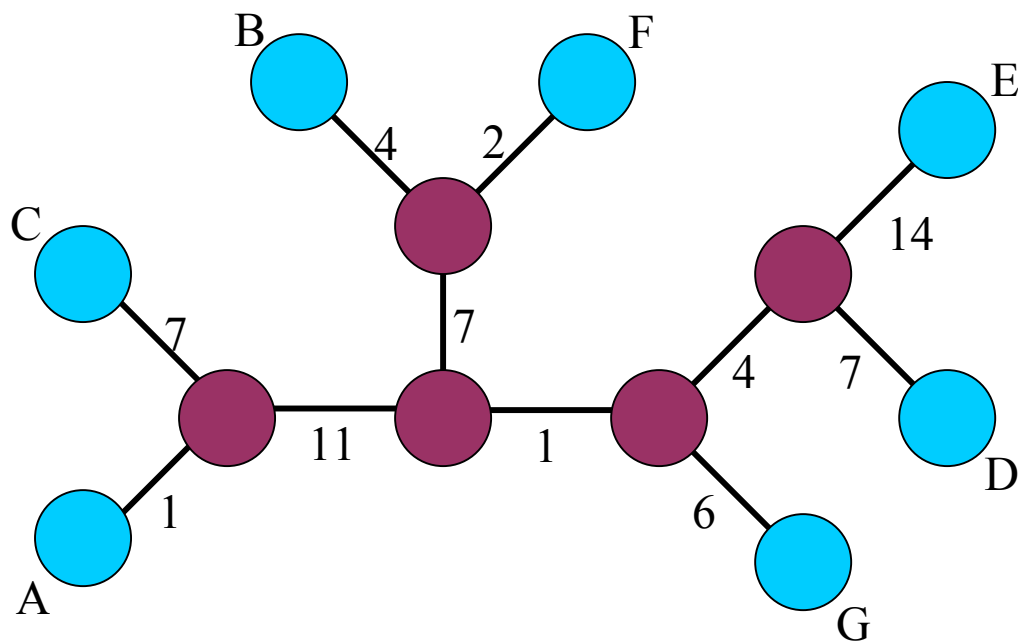
Distanzmatrix

	A	B	C	D	E	F	G
A	0	6	8	12	11	13	15
B	6	8	6	12	11	13	15
C	8	6	0	14	13	15	17
D	12	12	14	0	3	9	11
E	11	11	13	3	0	8	10
F	13	13	15	9	8	0	6
G	15	15	17	11	10	6	0

Wir wissen nicht, welche der Spezies welche **gemeinsamen Vorfahren** hatten, welche **Vorfahren** es überhaupt gab, und wie all diese sich **evolutionär auseinander entwickelt** haben. Was wird gerne hätten, wäre ein Baum,

- an dessen Blättern die sieben Spezies A – G stehen,
- an dessen inneren Knoten frühere Spezies stehen mit jeweils drei Nachbarknoten (1 Vater, 2 Söhne),
- und an dessen Kanten Evolutionsdauern stehen, so dass die Summe der Evolutionsdauern an den Kanten von

Blatt X zu Blatt Y gerade die Distanz zwischen X und Y aus der obigen Tabelle ist. Ein solcher Baum heißt dann **additiver Baum** zur gegebenen Distanzmatrix.



Es wäre $d(A,C) = 8$ in diesem Baum korrekt repräsentiert, $d(A,E) = 11$ dagegen nicht, da die „Pfadzeit“ im Baum von A bis E gleich $1 + 11 + 1 + 4 + 14 = 31$ ist.

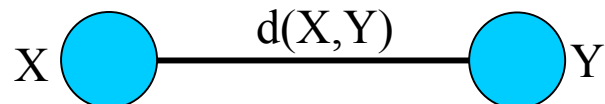
- Gibt es für eine Distanzmatrix immer einen additiven Baum?
- Kann es mehrere verschiedene geben?
- Ist das Modell des additiven Baumes überhaupt biologisch adäquat?

Frage: Wie viele binäre Bäume mit n Blättern gibt es (ohne Berücksichtigung von Kantenbeschriftungen)?

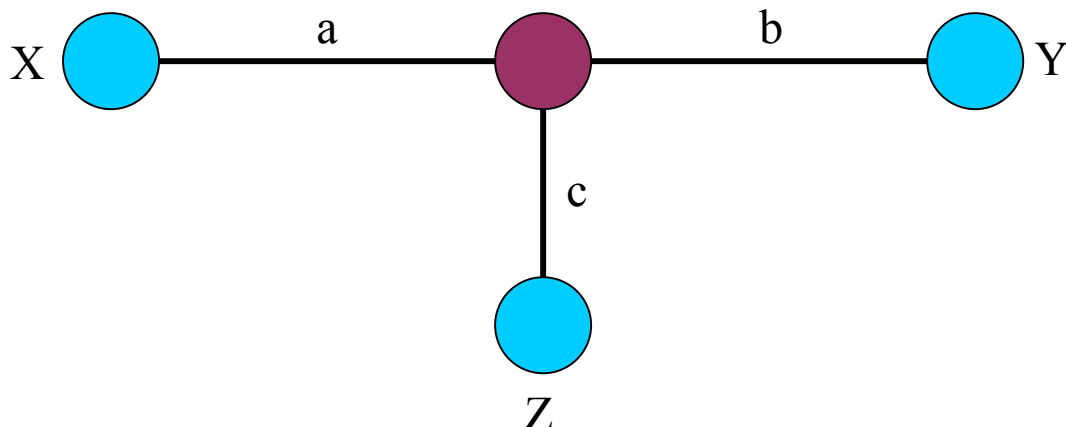
Die Antwort auf diese Frage zeigt, dass das systematische Durchsuchen aller möglichen binären Bäume unmöglich ist.

Fall 1: Zwei Spezies X und Y mit Distanz $d(X,Y)$.

Der eindeutig bestimmte additive Baum ist:



Fall 2: Drei Spezies X, Y und Z mit Distanzen $d(X,Y)$, $d(X,Z)$ und $d(Y,Z)$. Ein additiver Baum kann nur wie folgt aussehen, mit zunächst noch unbekannten Evolutionszeiten a, b, c :



Es müssen gelten:

$$d(X,Y) = a + b \quad d(X,Z) = a + c \quad d(Y,Z) = b + c$$

Wir bestimmen die Evolutionszeiten daraus:

$$d(X,Y) - d(X,Z) = b - c$$

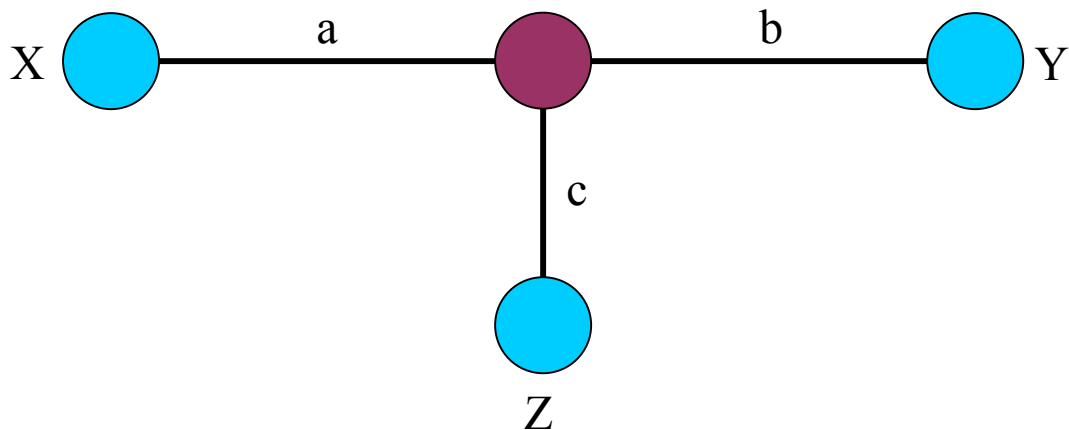
$$d(X,Y) - d(X,Z) + d(Y,Z) = 2b$$

$$b = \frac{1}{2} (d(X,Y) + d(Y,Z) - d(X,Z))$$

$$\begin{aligned} a &= d(X,Y) - b \\ &= d(X,Y) - \frac{1}{2} (d(X,Y) + d(Y,Z) - d(X,Z)) \\ &= \frac{1}{2} (d(X,Y) + d(X,Z) - d(Y,Z)) \end{aligned}$$

$$\begin{aligned} c &= d(X,Z) - a \\ &= d(X,Z) - \frac{1}{2} (d(X,Y) + d(X,Z) - d(Y,Z)) \\ &= \frac{1}{2} (d(X,Z) + d(Y,Z) - d(X,Y)) \end{aligned}$$

- In der Zusammenfassung und mit einem Blick auf den Baum wird das „Berechnungsmuster“ sichtbar:



$$a = \frac{1}{2} (d(X,Y) + d(X,Z) - d(Y,Z))$$

$$b = \frac{1}{2} (d(X,Y) + d(Y,Z) - d(X,Z))$$

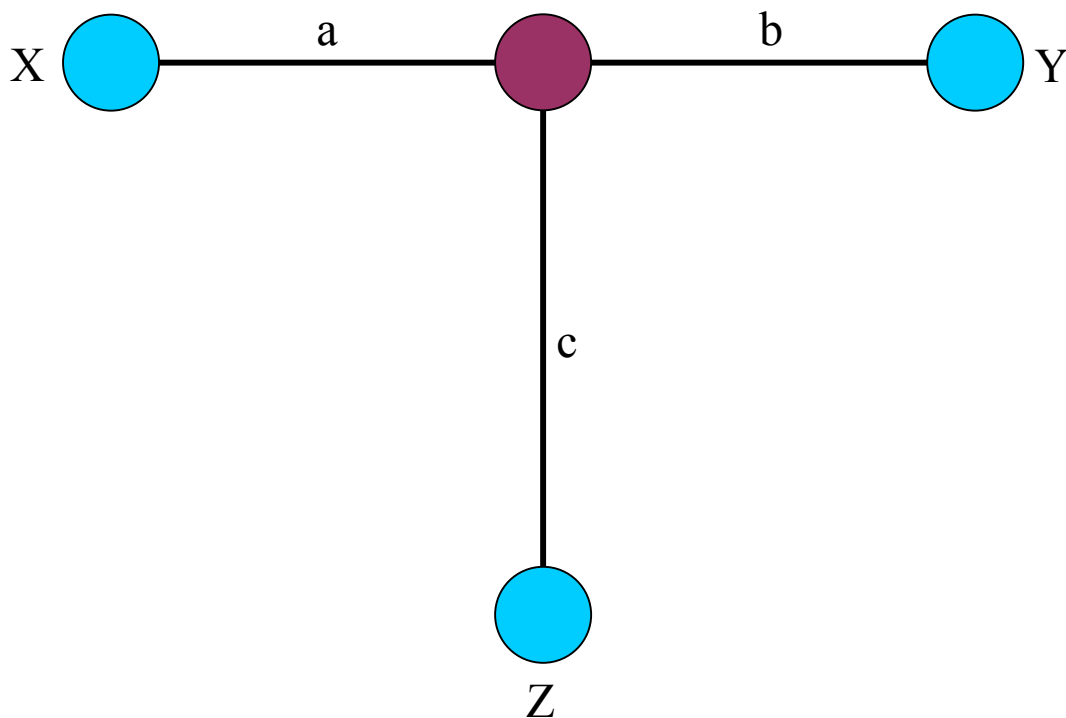
$$c = \frac{1}{2} (d(X,Z) + d(Y,Z) - d(X,Y))$$

- Wer garantiert uns, dass die Evolutionszeiten a, b und nichtnegativ sind? Die **Dreiecksungleichung** garantiert dies – vorausgesetzt, sie gilt: $d(X,Y) \leq d(X,Z) + d(Z,Y)$
- Wer garantiert uns, dass die Evolutionszeiten a, b und nicht gleich Null sind? Niemand. Ist eine der drei Zeiten gleich 0, so gibt es eben keinen additiven Baum für die drei Objekte.

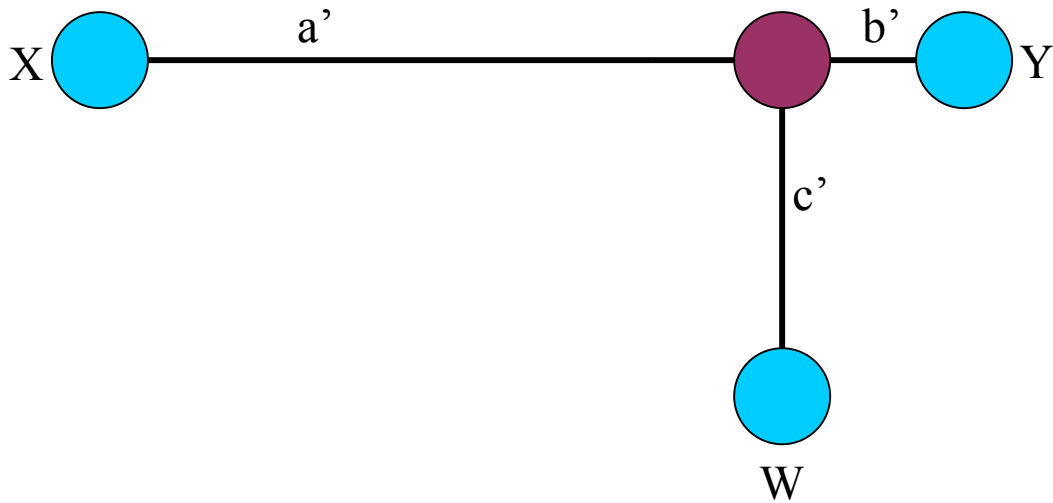
Fazit: Zwei beliebig gewählte Referenzspezies X und Y legen eindeutig fest, wie relativ zu ihnen eine weitere Spezies Z in einen möglichen additiven Baum einzufügen ist.

Fall 3: Vier Spezies X, Y, Z und W mit Distanzen $d(X,Y)$, $d(X,Z)$, $d(X,W)$, $d(Y,Z)$, $d(Y,W)$ und $d(Z,W)$.

Wir konstruieren zunächst wie in Fall 2 den eindeutig bestimmten (wenn überhaupt einer existiert) additiven Baum für X, Y und Z:

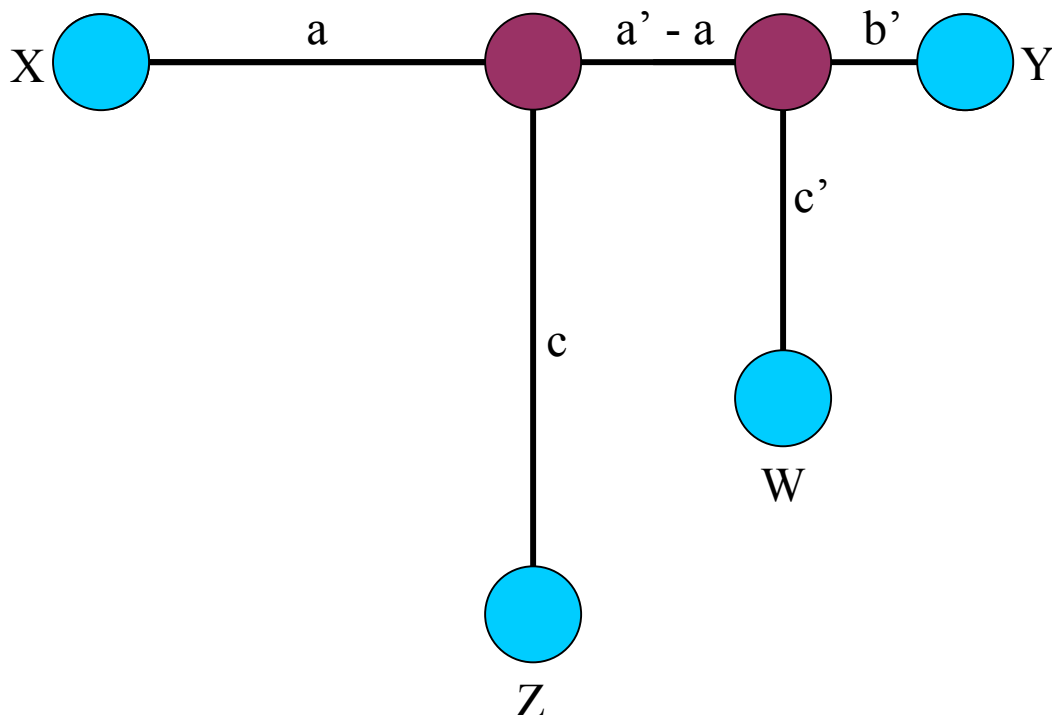


Nun benutzen wir beispielsweise X und Y als Referenzspezies, um die Platzierung von W in höchstens eindeutiger Weise vorzunehmen.



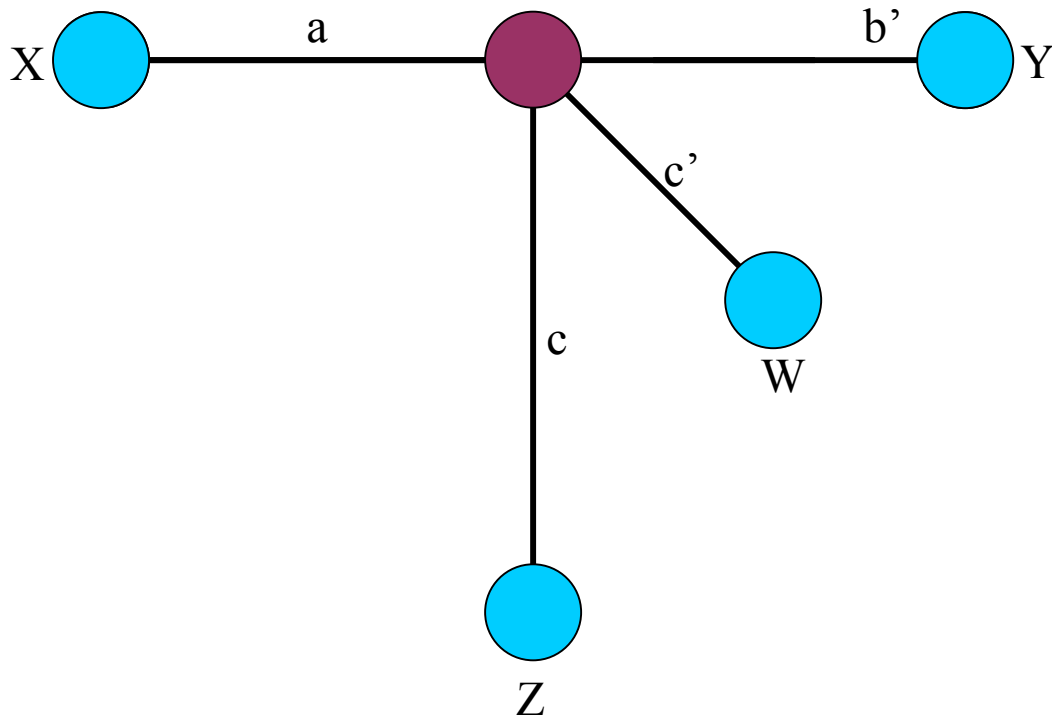
Dabei können zwei Fälle eintreten:

Fall 3a: Die beiden in den beiden oben konstruierten Bäumen auftretenden Vorfahren zwischen X und Y sind verschiedenen, d.h. $a \neq a'$. Somit ist die einzige Chance auf einen additiven Gesamtbaum die folgende:



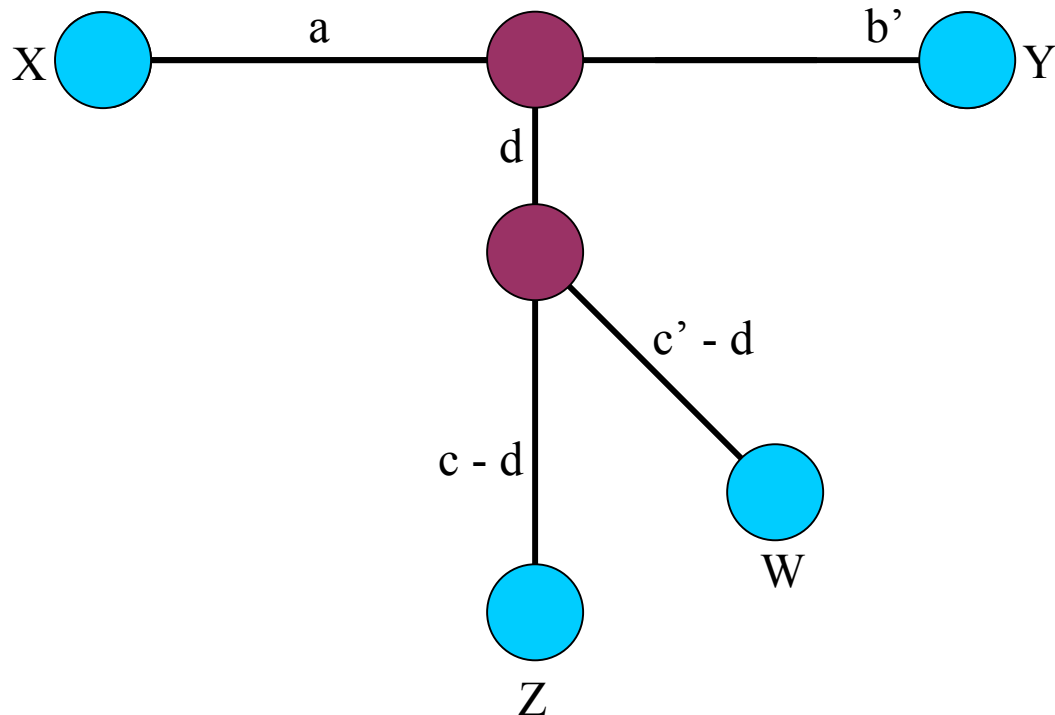
Nun müssen wir nur noch überprüfen, ob die eindeutige Platzierung von W auch noch mit den von X und Y verschiedenen Knoten kompatibel ist (in diesem Fall also mit Z). Ist dies der Fall, haben wir den eindeutigen additiven Baum für alle vier Spezies gefunden, andernfalls wissen wir nun, dass es keinen solchen gibt.

Fall 3b: Die beiden in den beiden oben konstruierten Bäumen auftretenden Vorfahren zwischen X und Y sind identisch, d.h. $a = a'$. Das geht in unserem Modell nicht!



Noch aber sind wir nicht am Ende! Es sei etwa $c' < c$ (bei $c < c'$ geht es genauso, bei $c = c'$ haben wir keinen Ausweg mehr). Dann können wir versuchen, auf der Kante mit Wert c einen weiteren Vorfahren zu platzieren, der dann zu Nachkommen Z und W führt. Allerdings müssen die Evolutionszeiten wieder passen. Wie dies geht, bekommen wir heraus, wenn wir versuchen, W mit Referenzspezies X und Z (statt wie vorhin X und Y) zu platzieren.

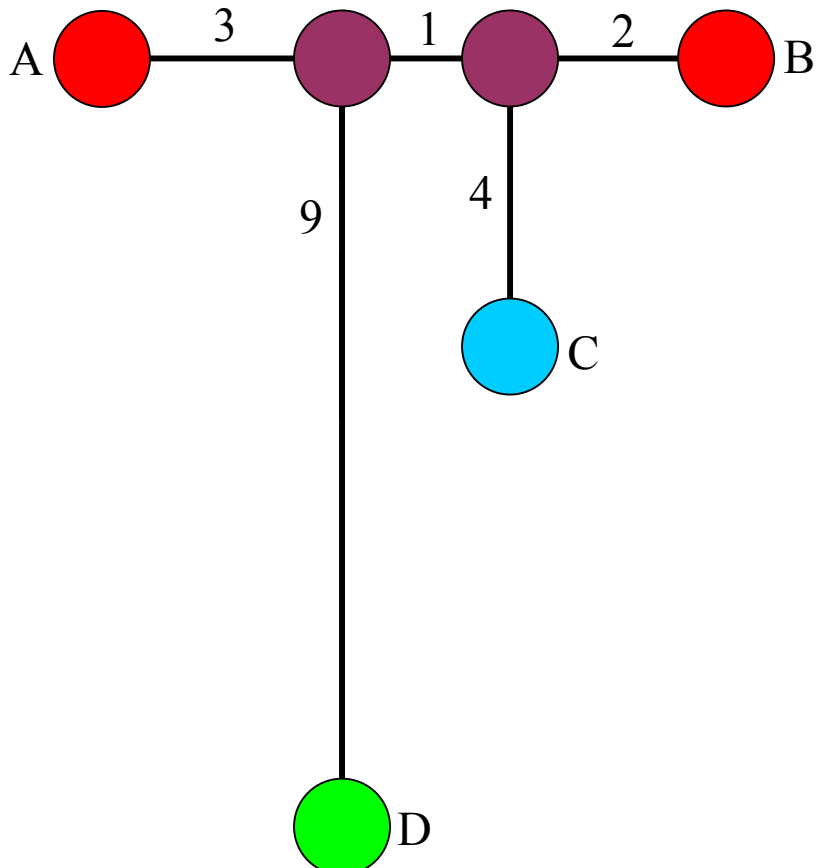
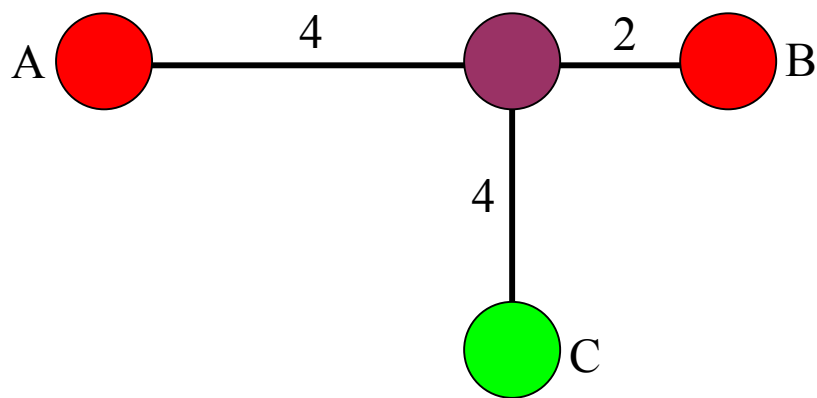
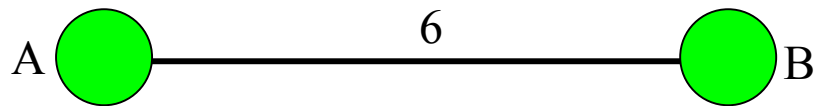
Wenn wir Glück haben, ergibt sich ein additiver Baum folgender Gestalt, in dem die Neuplatzierung von W auch noch mit allen anderen Knoten kompatibel ist:

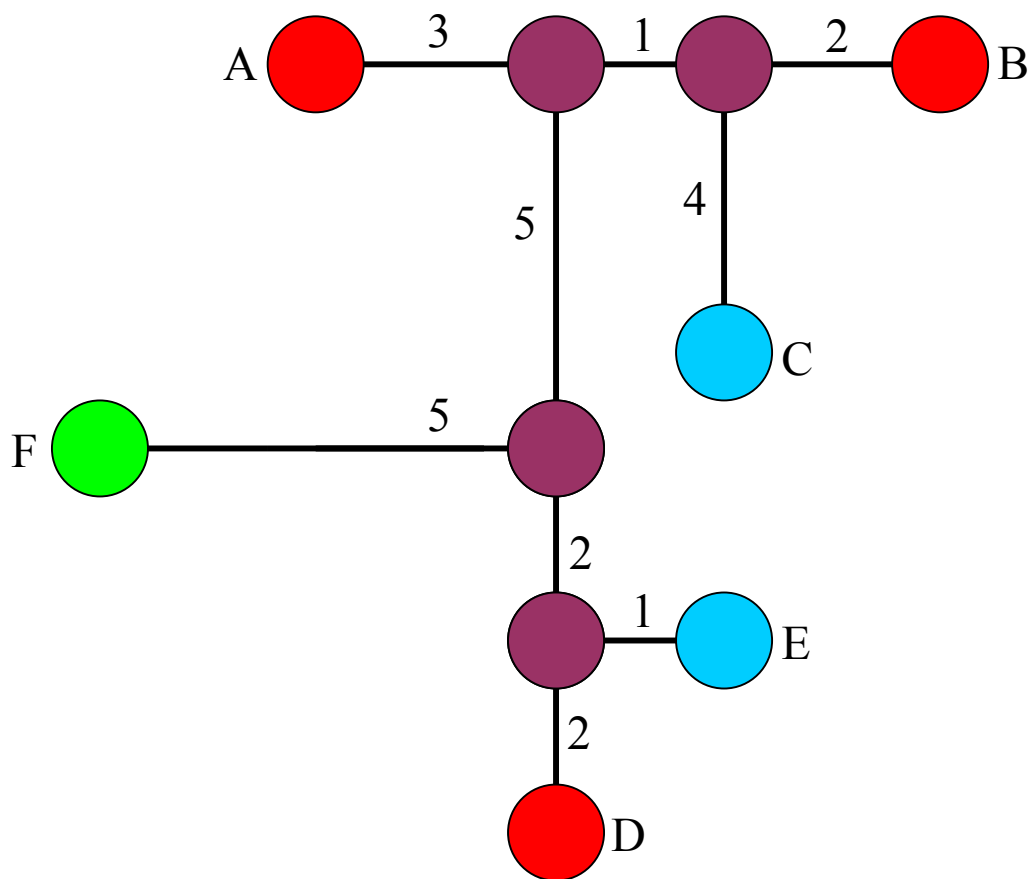
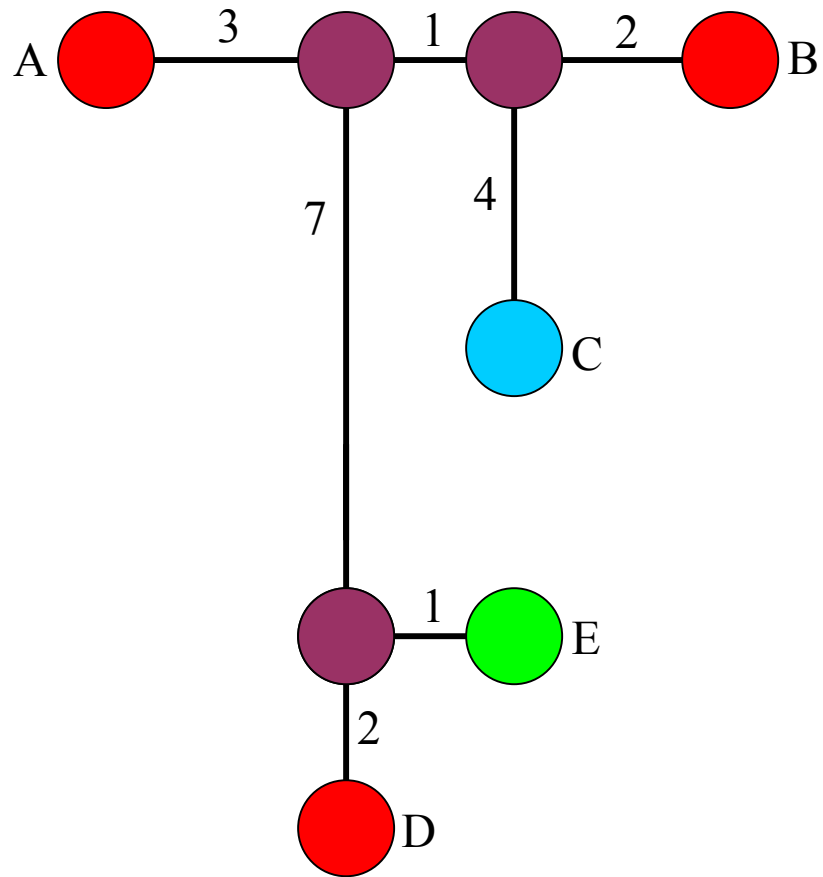


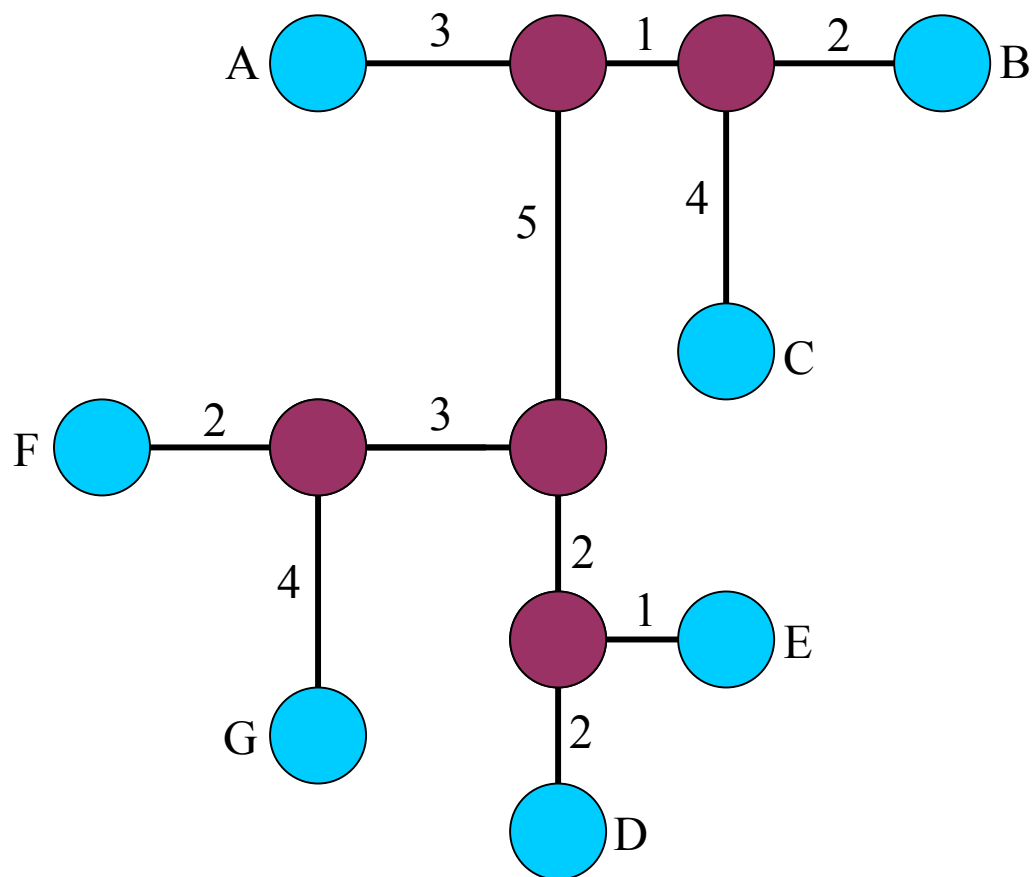
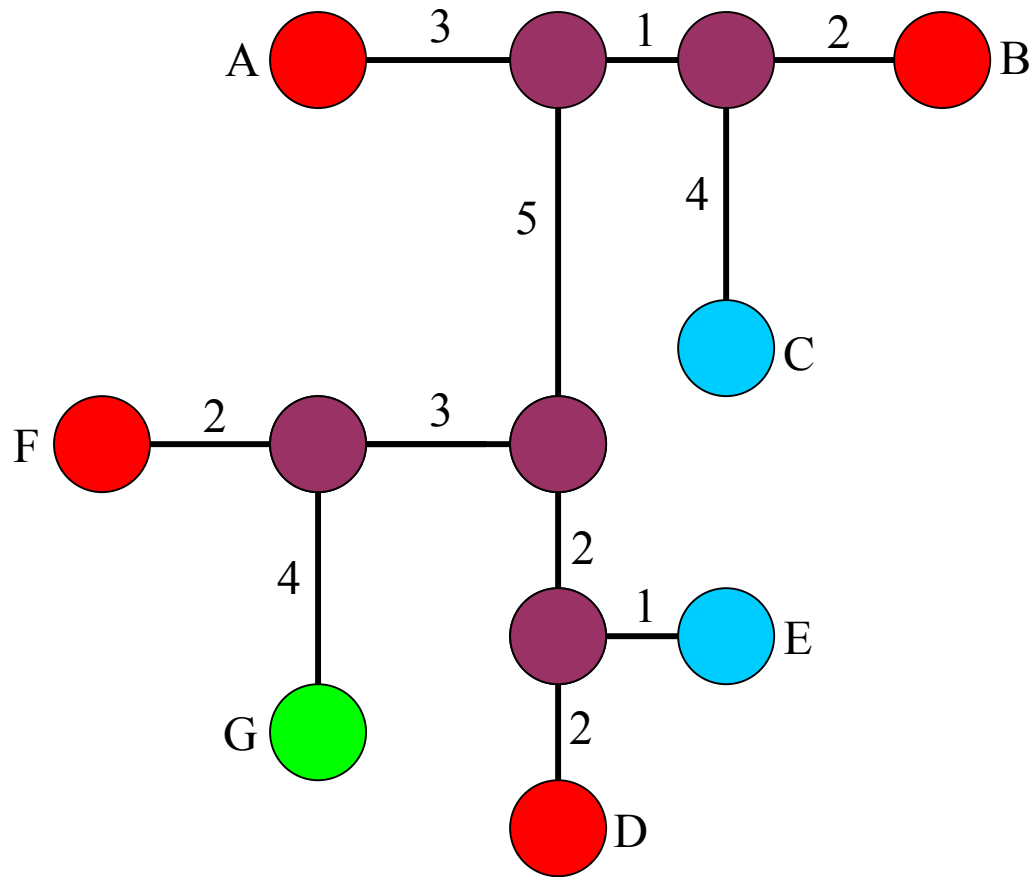
Fazit: Mit maximal zwei Versuchen, einen weiteren neuen Knoten passend in einen bereits partiell aufgebauten additiven Baum zu platzieren, können wir einen additiven Baum um eine weitere Spezies zu einem wiederum additiven Baum erweitern, oder feststellen, dass kein solcher existiert.

Dieses Verfahren sollten wir einmal für unsere Phantasiegestalten abrollen lassen:

Die jeweils benutzen Referenzspezies sind farblich **rot** markiert; der jeweils neue Punkt ist **grün** markiert. Die jeweiligen Berechnungen sind nicht aufgeführt (bitte nachrechnen):





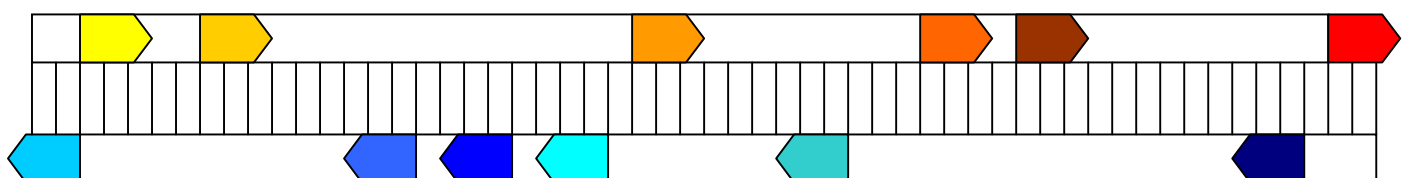


Thema 2: Genom-Rearrangement

Woher kommen denn nun die Distanzmatrizen, die wie bei der Konstruktion von Stammbäumen verwendet haben?

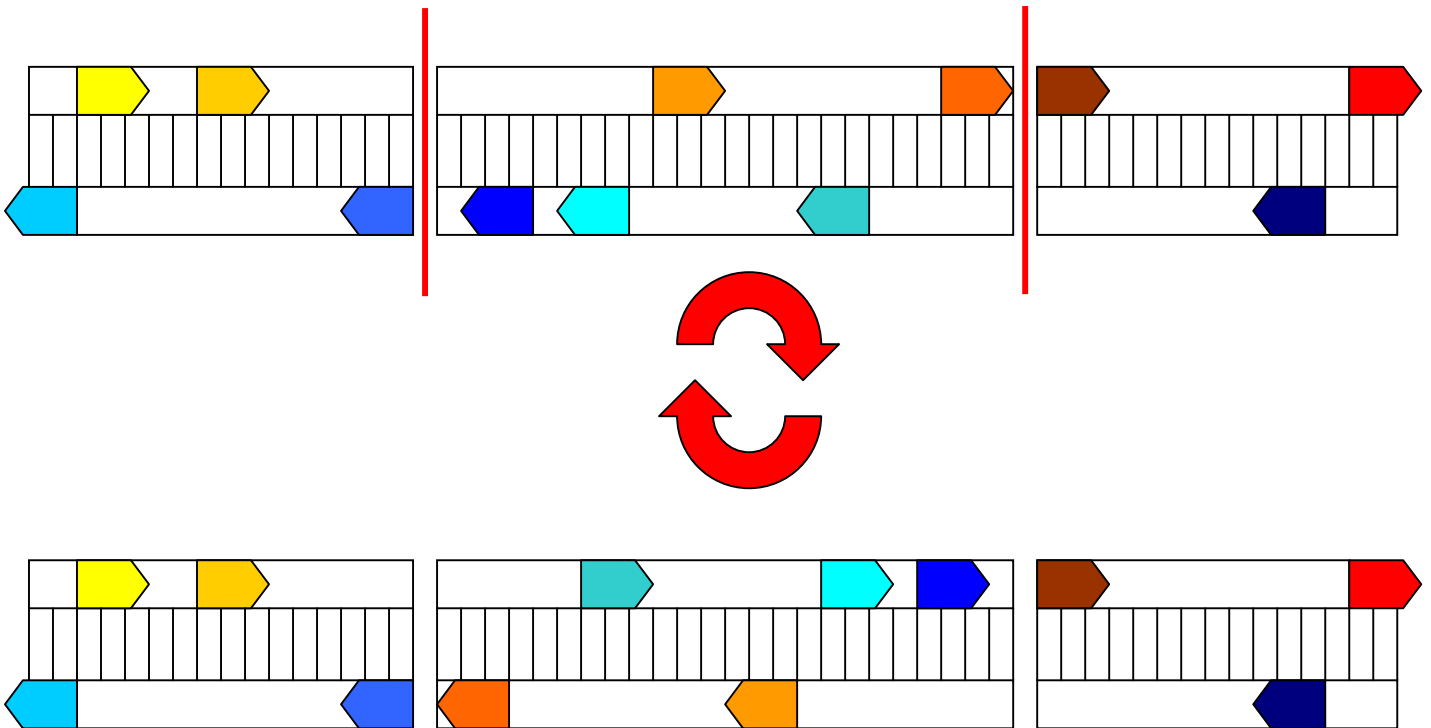
Beispielsweise aus der Analyse der **Reihenfolge von Genen** auf den Chromosomen.

Ein Gen ist ein Chromosomenabschnitt, der die Herstellung eines Proteins bewirkt oder ein Signal für einen biochemischen Prozess darstellt; ein Gen kann auf dem einen oder anderen Strang der Doppelhelix liegen. Stränge sind biochemisch mit einer Richtung versehen. Wir modellieren dies durch ein Vorzeichen + („oberer Strang“) oder – („unterer Strang“):



-1 +2 +3 -4 -5 -6 +7 -8 +9 +10 -11 +12

Relativ selten, aber in ziemlich konstanten Zeitintervallen,
vollzieht die Evolution „gerichtete Umordnungen“:



-1 +2 +3 -4 -5 -6 +7 -8 +9 +10 -11 +12

-1 +2 +3 -4 -9 +8 -7 +6 +5 +10 -11 +12

Wir wollen das Chromosom mit der Genreihenfolge

-1 +2 +3 -4 -5 -6 +7 -8 +9 +10 -11 +12

transformieren in das Chromosom mit der Genreihenfolge

-2 +3 -4 -12 +5 -10 -9 +8 -1 -11 -6 +7

-1 +2 +3 -4 -5 -6 +7 -8 +9 +10 -11 +12

-1 +2 +3 -4 -9 +8 -7 +6 +5 +10 -11 +12

-1 -6 +7 -8 +9 +4 -3 -2 +5 +10 -11 +12

-1 -6 +7 -8 +9 +4 -5 +2 +3 +10 -11 +12

-1 -6 +7 -8 +9 +4 -5 +2 +3 +10 +11 +12

-1 -2 +5 -4 -9 +8 -7 +6 +3 +10 +11 +12

-1 -2 +5 -4 -3 -6 +7 -8 +9 +10 +11 +12

-1 -2 +3 +4 +5 -6 +7 -8 +9 +10 +11 +12

-7 +6 -5 -4 -3 +2 +1 -8 +9 +10 +11 +12

-7 +6 -5 -4 -11 -10 -9 +8 -1 -2 +3 +12

-7 +6 -5 -4 -11 -10 -9 +8 -1 -12 -3 +2

-7 +6 +11 +4 +5 -10 -9 +8 -1 -12 -3 +2

-7 +6 +11 +1 -8 +9 +10 -5 -4 -12 -3 +2

-7 +6 +11 +1 -8 +9 +10 -5 +12 +4 -3 +2

-2 +3 -4 -12 +5 -10 -9 +8 -1 -11 -6 +7

Frage: Geht es mit weniger als 14 Genom-Umordnungen?
Was ist die minimale Anzahl von Genom-Umordnungen in diesem Beispiel?

Antwort: Es geht mit 7 Umordnungen, aber nicht mit weniger.

Schreibe Startchromosom

-1 +2 +3 -4 -5 -6 +7 -8 +9 +10 -11 +12

als Abfolge von „Batterien“ mit Plus- und Minus-Polen:

$[-1+1] [+2-2] [+3-3] [-4+4] [-5+5] [+7-7] [-8+8] [+9-9]$
 $[+10-10] [-11+11] [+12-12]$

Ebenso das Zielchromosom

-2 +3 -4 -12 +5 -10 -9 +8 -1 -11 -6 +7

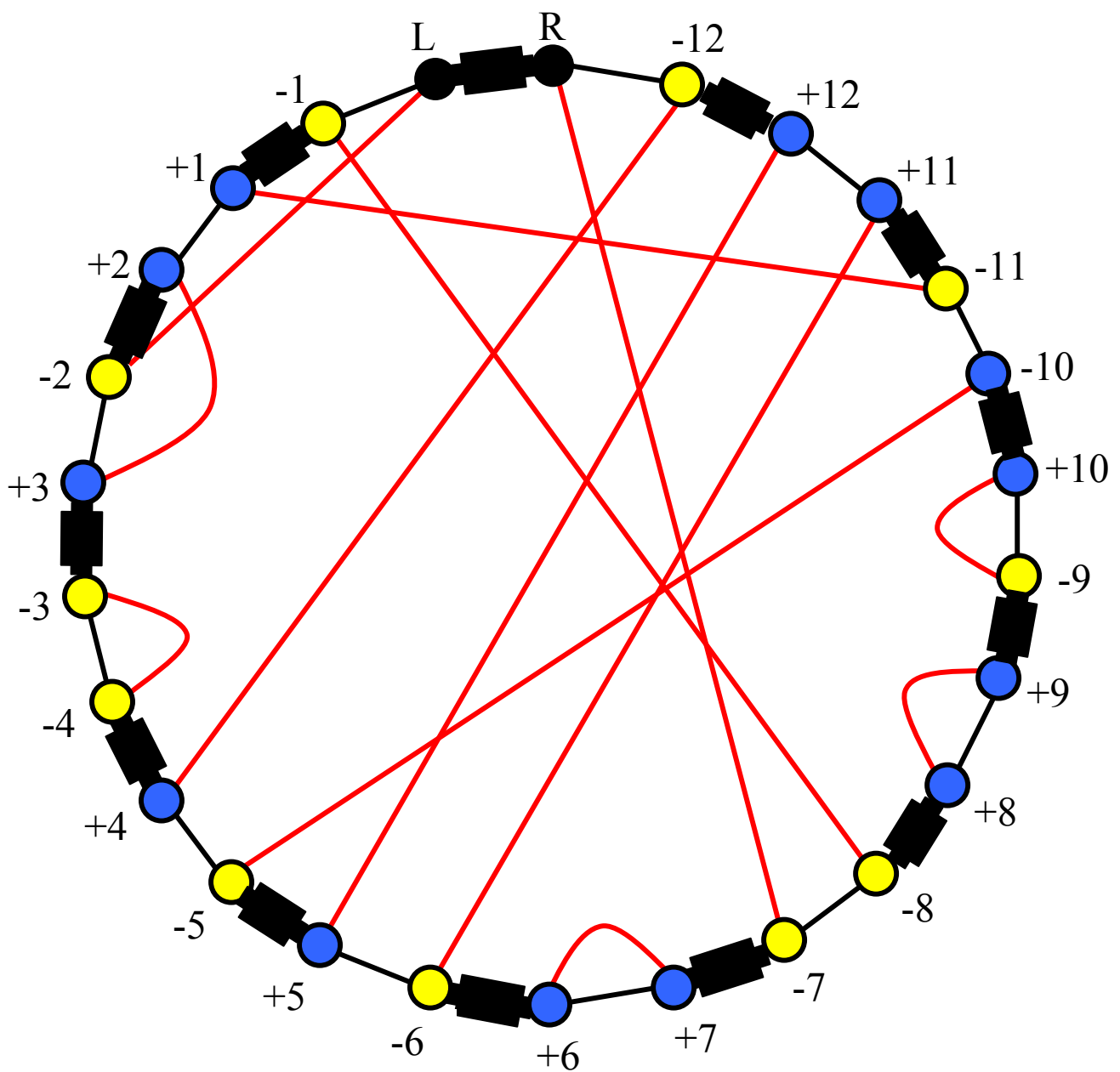
$[-2+2] [+3-3] [-4+4] [-12+12] [+5-5] [-10+10] [-9+9]$
 $[+8-8] [-1+1] [-11+11] [-6+6] [+7-7]$

Das Reality-Desire-Diagramm:

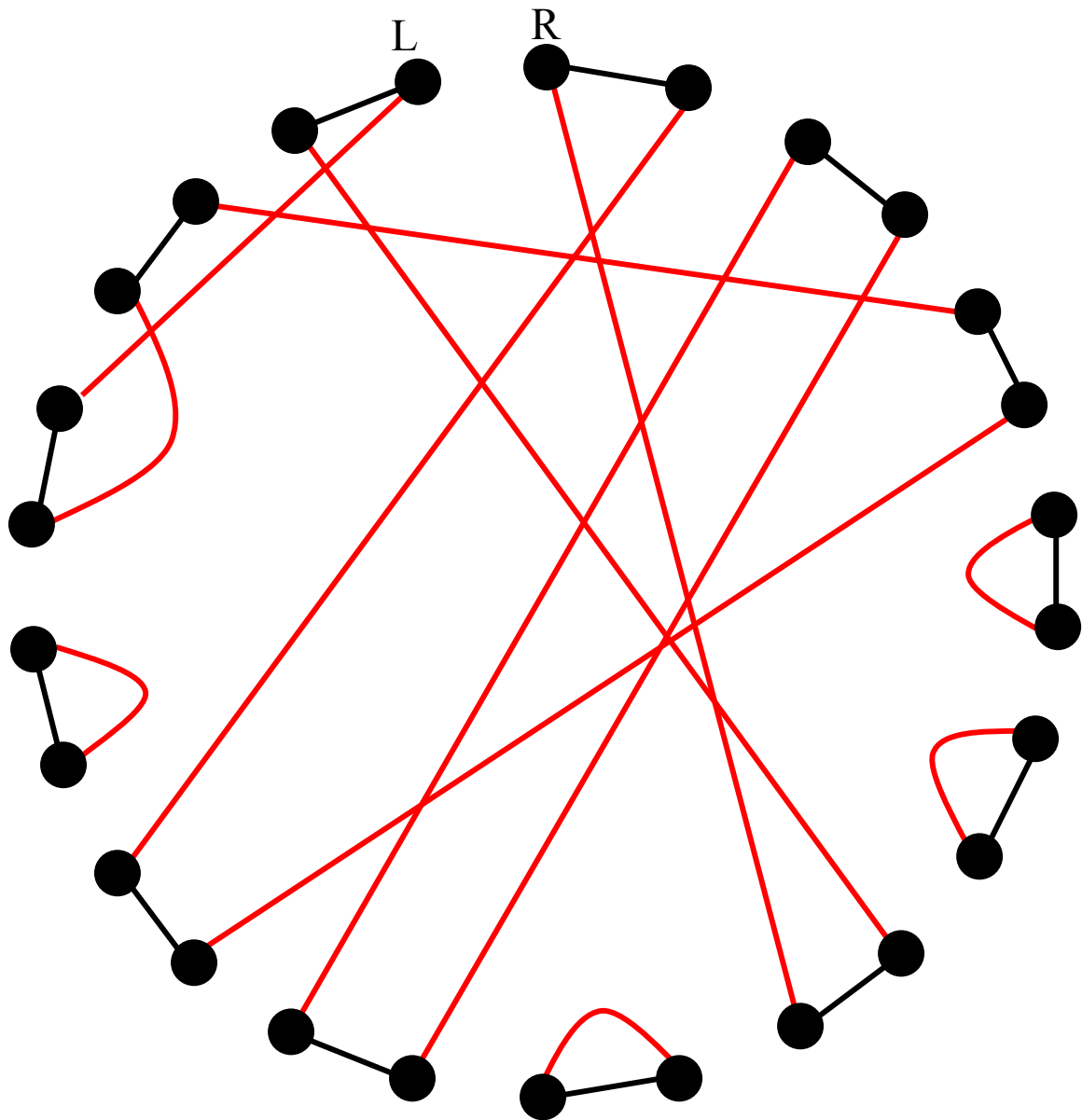
Gene = „Batterien mit Polung“

„Drähte“ auf dem Kreisrand = Reality (so ist es)

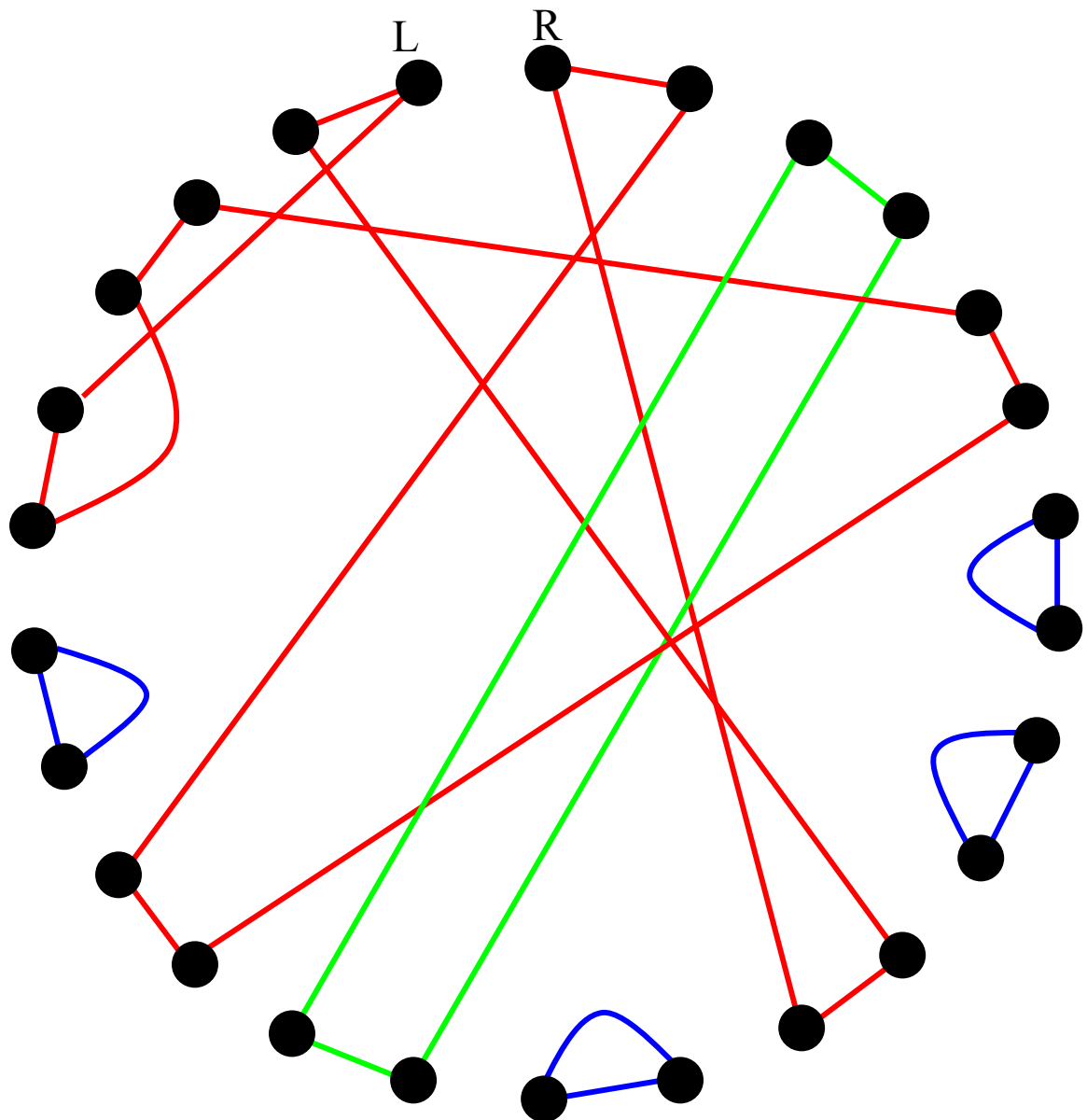
„Drähte“ im Kreisinneren = Desire (so soll es sein)



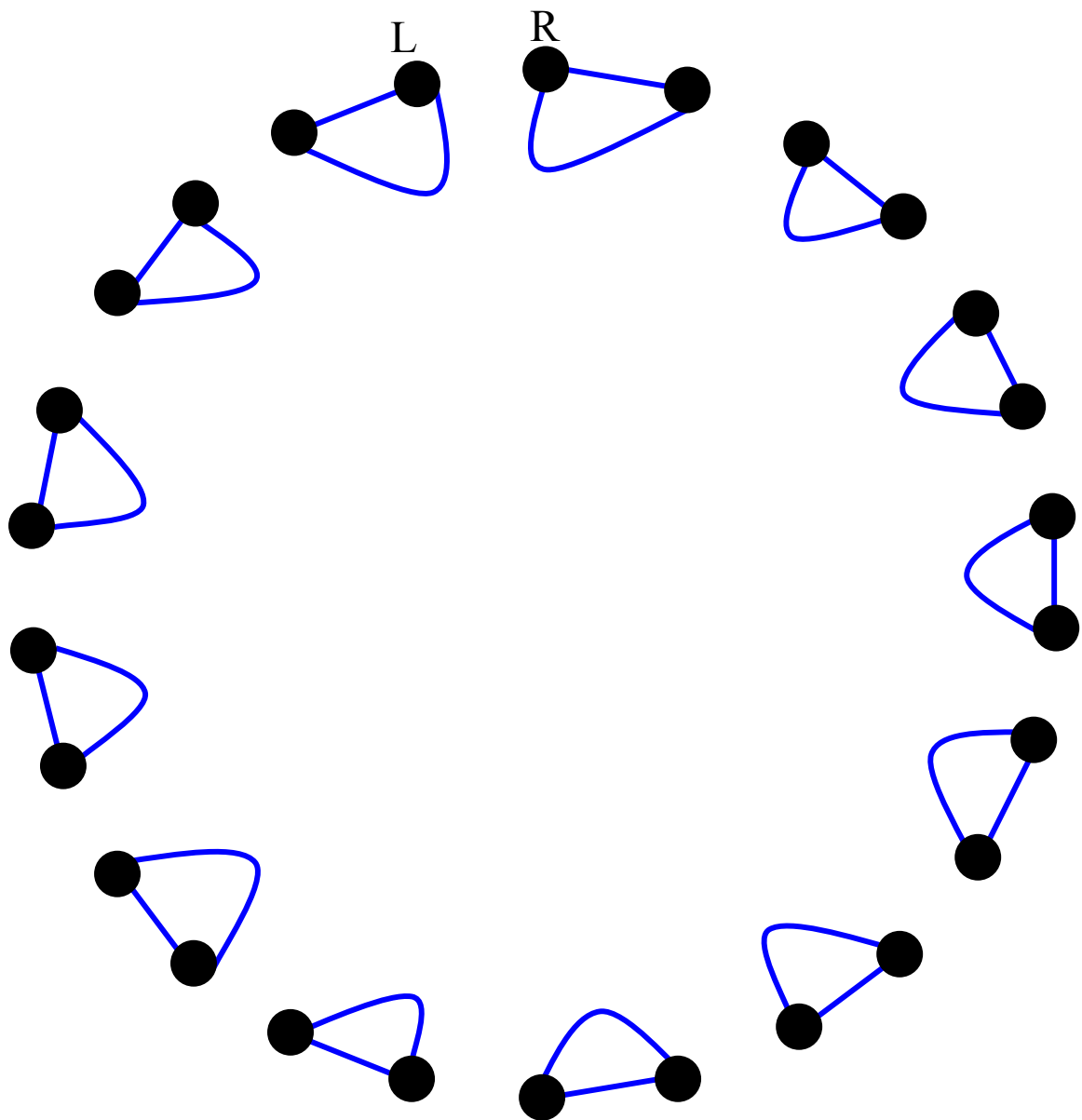
Weglassen unwesentlicher Details



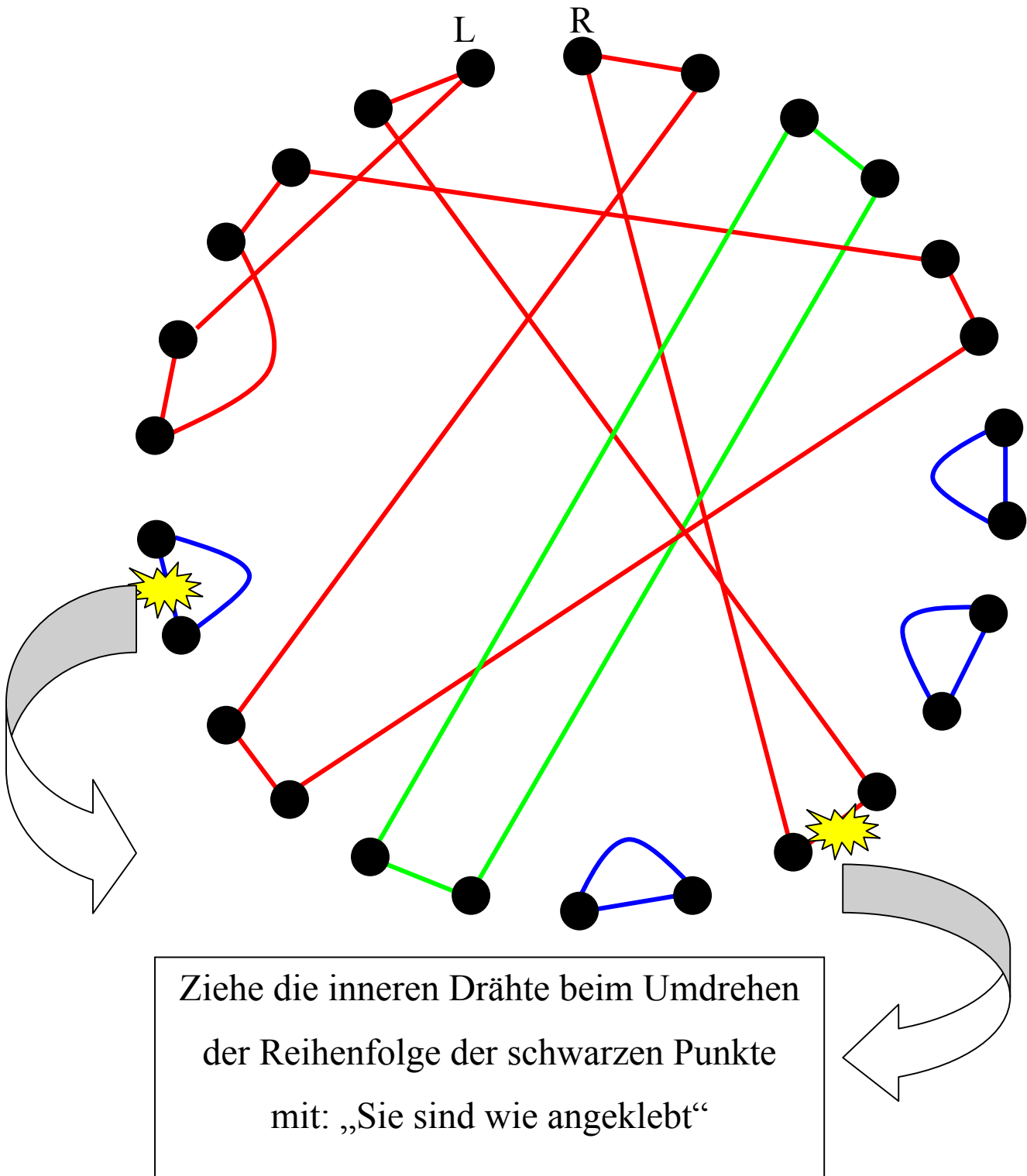
Kenntlich machen wichtiger Strukturelemente: 6 Kreise



Zieldiagramm hat 13 triviale Kreise

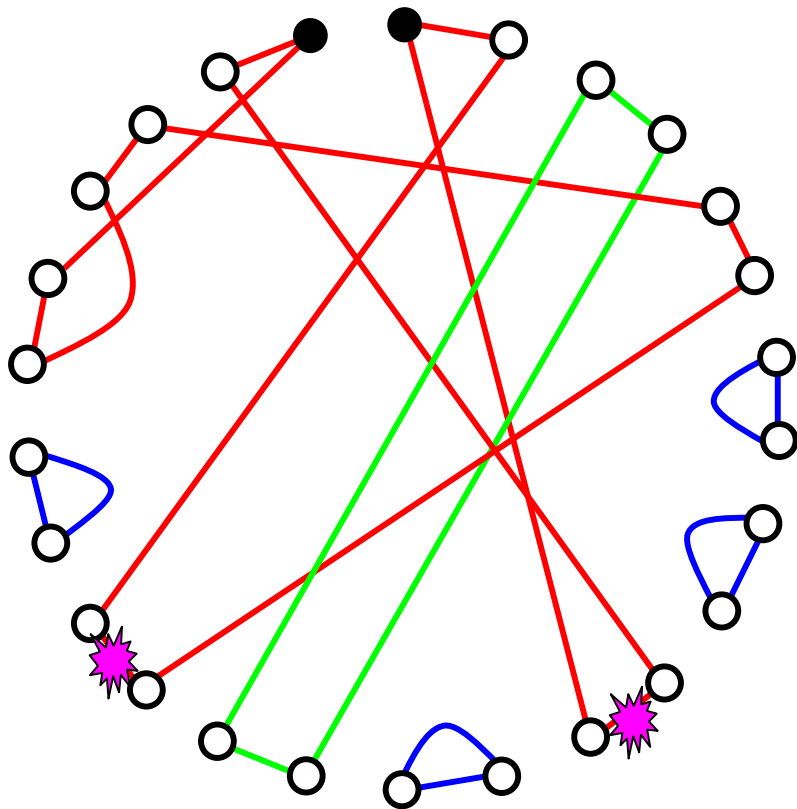


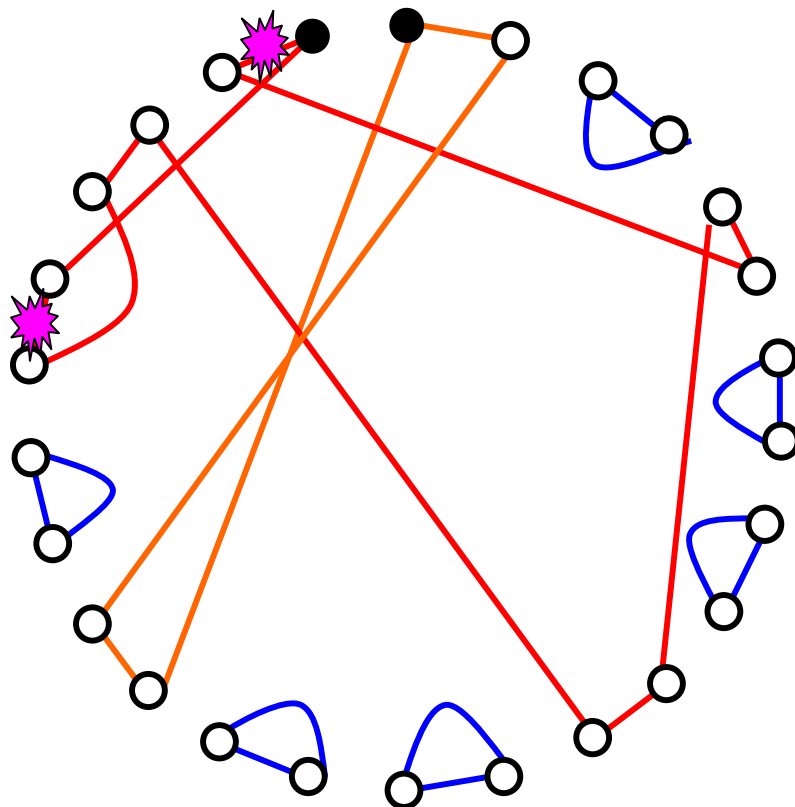
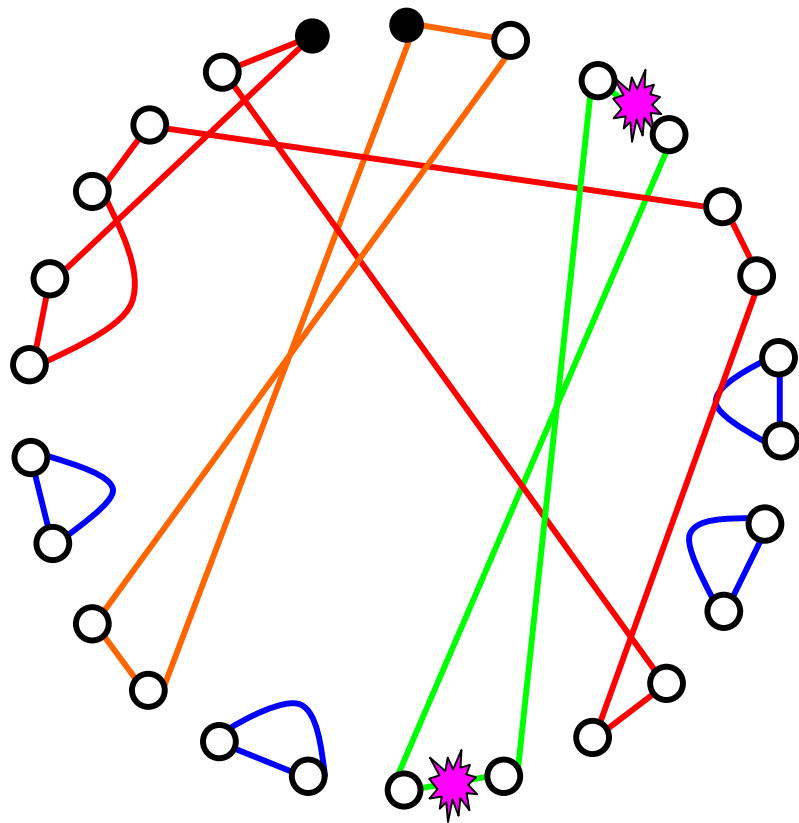
Effekt einer Umordnung im Reality-Desire-Diagramm

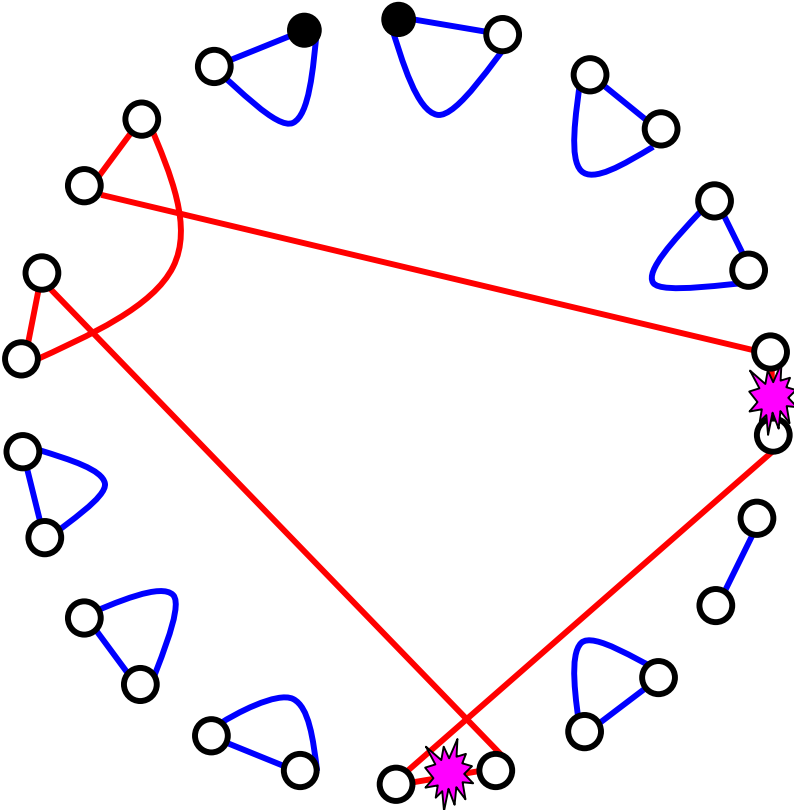
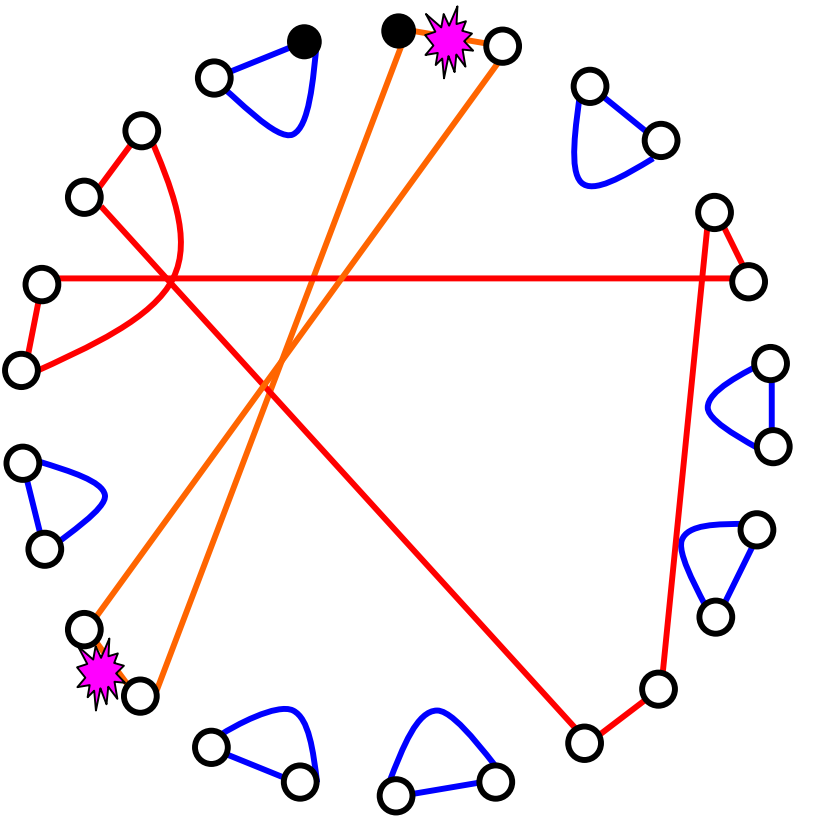


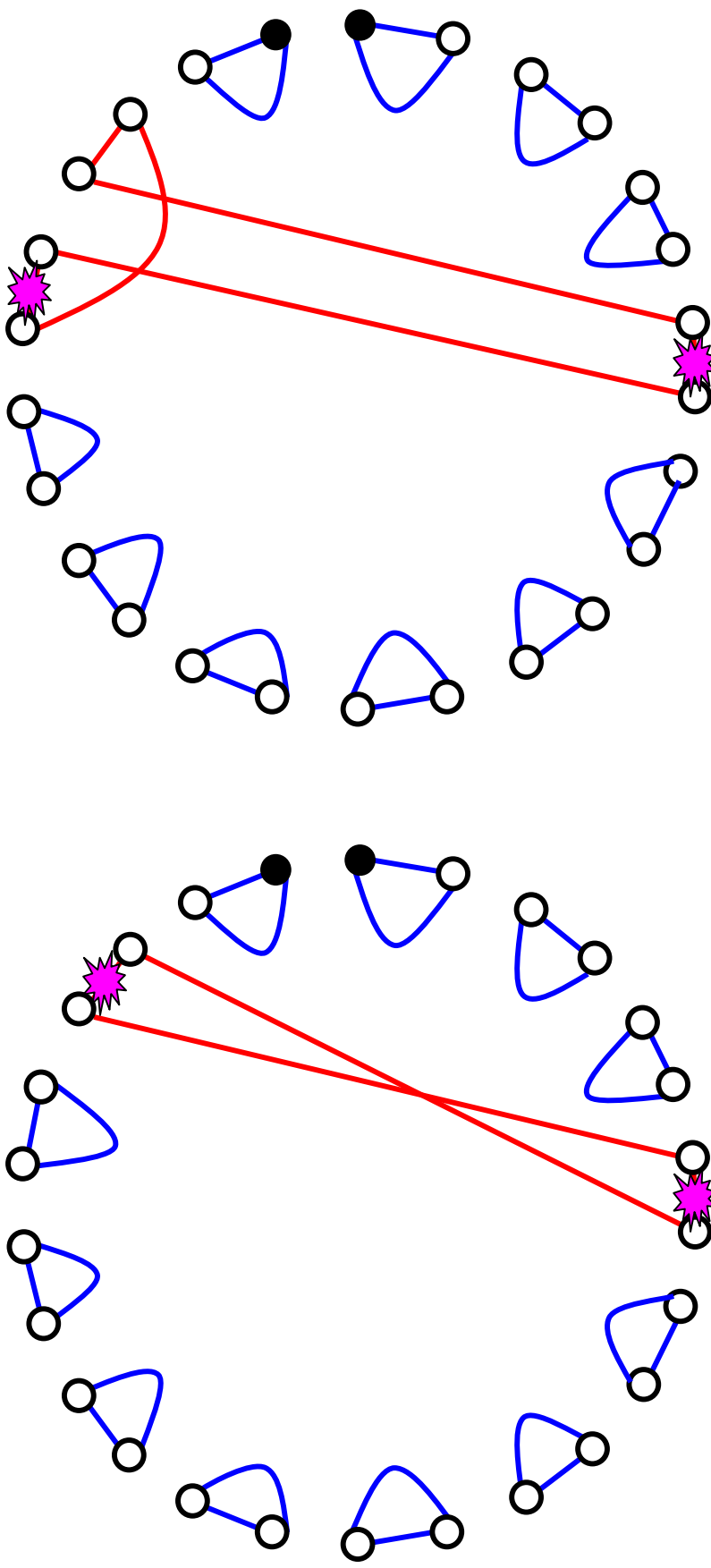
Satz: Bei einer Umordnung kann die Zahl der Kreise um 1 wachsen („hurra“), gleich bleiben („schade“) oder um 1 sinken („oh weh“).

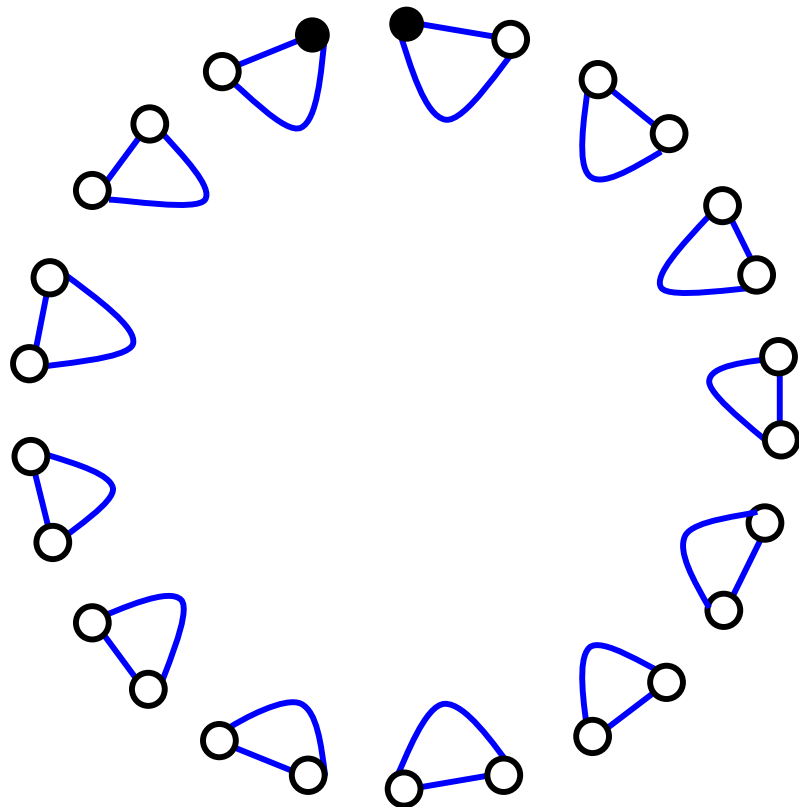
Wie man durch 7 geeignete Umordnungen immer mehr Kreise erzeugt, bis hin zu 13 trivialen Kreisen:











Prinzip: Ordne möglichst an zwei gegenläufig orientierten Kanten eines einzigsten Kreises um. Dann steigt die Anzahl der Kreise auf jeden Fall um 1.

Bei zwei gleichläufig orientierten Kanten eines einzigen Kreises bleibt die Anzahl der Kreise unverändert.

Bei zwei Kanten in verschiedenen Kreises sinkt die Anzahl der Kreise um 1.

So einfach geht es nicht immer. Es gibt nicht immer solche Kanten. Es können auch gegenläufig orientierte Kanten gleichläufig werden.

Mathematische Analyse all dieser Problemfälle ergibt:

Satz Die minimale Anzahl erforderlicher gerichteter Umordnungen ist:

$$\mathbf{n + 1}$$

– Kreise

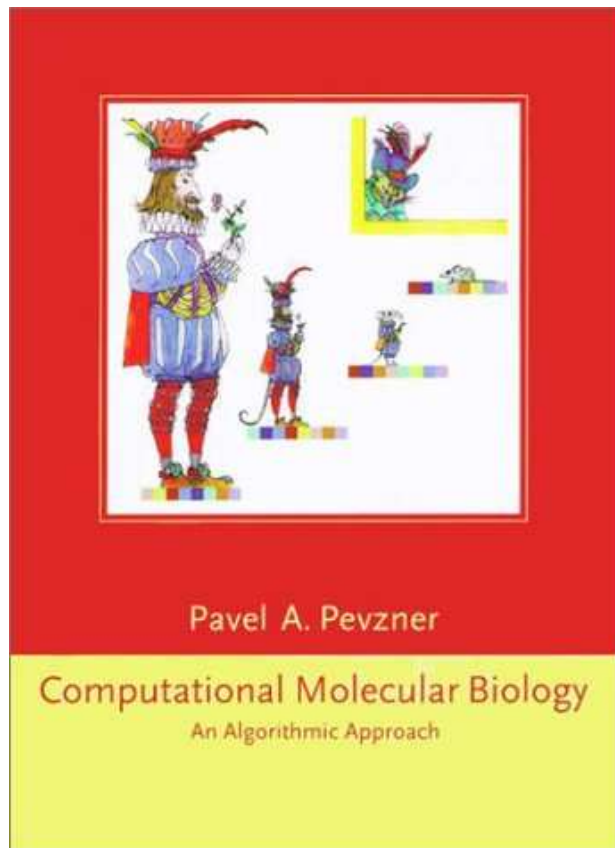
+ Hürden

+ Festungen

wobei wir die komplizierten graphischen Elemente Hürden und Festungen hier nicht definiert haben.

In unserem Beispiel ergibt sich $13 - 6 + 0 + 0 = 7$.

Ungerichtete Umordnungen – wenn die Genausrichtung unbekannt ist



Turning Man To Worm

36 Gene auf der Mitochondrien-DNA eines Wurmes werden durch eine Folge von 26 ungerichteten Reversals in die Reihenfolge gebracht werden, wie sie beim Menschen vorliegt

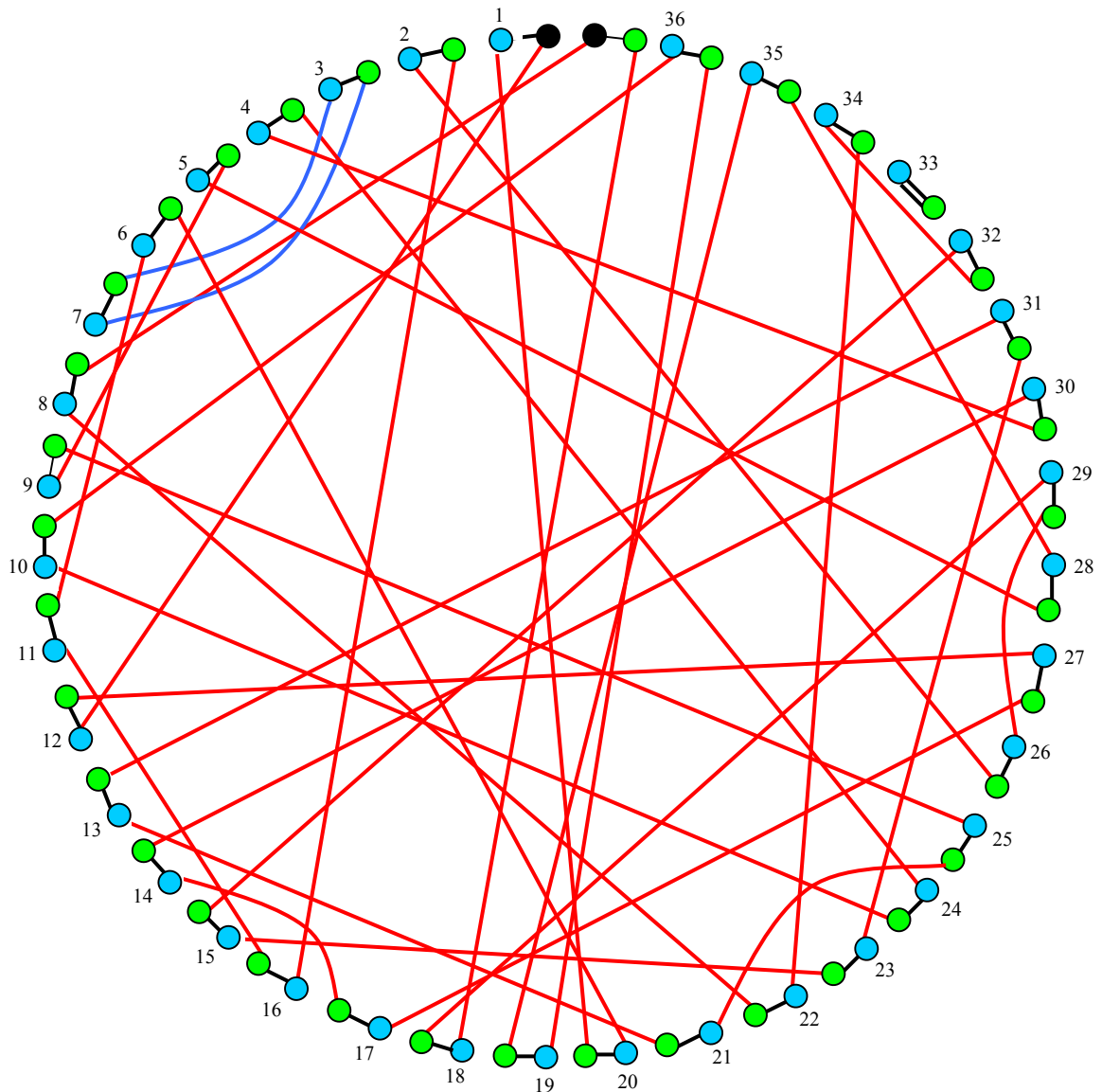
Bioinformatik

12 31 34 28 26 17 29 04 09 36 18 35 19 01 16 14 32 33 22 15 11 27 05 20 13 30 23 10 06 03 24 21 08 25 02 07
20 05 27 11 15 22 33 32 14 16 01 19 35 18 36 09 04 29 17 26 28 34 31 12 13 30 23 10 06 03 24 21 08 25 02 07
01 16 14 32 33 22 15 11 27 05 20 19 35 18 36 09 04 29 17 26 28 34 31 12 13 30 23 10 06 03 24 21 08 25 02 07
01 16 15 22 33 32 14 11 27 05 20 19 35 18 36 09 04 29 17 26 28 34 31 12 13 30 23 10 06 03 24 21 08 25 02 07
01 16 15 36 18 35 19 20 05 27 11 14 32 33 22 09 04 29 17 26 28 34 31 12 13 30 23 10 06 03 24 21 08 25 02 07
01 16 15 14 11 27 05 20 19 35 18 36 32 33 22 09 04 29 17 26 28 34 31 12 13 30 23 10 06 03 24 21 08 25 02 07
01 16 15 14 31 34 28 26 17 29 04 09 22 33 32 36 18 35 19 20 05 27 11 12 13 30 23 10 06 03 24 21 08 25 02 07
01 26 28 34 31 14 15 16 17 29 04 09 22 33 32 36 18 35 19 20 05 27 11 12 13 30 23 10 06 03 24 21 08 25 02 07
01 26 28 18 36 32 33 22 09 04 29 17 16 15 14 31 34 35 19 20 05 27 11 12 13 30 23 10 06 03 24 21 08 25 02 07
01 26 28 29 04 09 22 33 32 36 18 17 16 15 14 31 34 35 19 20 05 27 11 12 13 30 23 10 06 03 24 21 08 25 02 07
01 26 28 29 30 13 12 11 27 05 20 19 35 34 31 14 15 16 17 18 36 32 33 22 09 04 23 10 06 03 24 21 08 25 02 07
01 26 11 12 13 30 29 28 27 05 20 19 35 34 31 14 15 16 17 18 36 32 33 22 09 04 23 10 06 03 24 21 08 25 02 07
01 26 27 28 29 30 13 12 11 05 20 19 35 34 31 14 15 16 17 18 36 32 33 22 09 04 23 10 06 03 24 21 08 25 02 07
01 26 27 28 29 30 31 34 35 19 20 05 11 12 13 14 15 16 17 18 36 32 33 22 09 04 23 10 06 03 24 21 08 25 02 07
01 26 27 28 29 30 31 34 35 19 20 09 22 33 32 36 18 17 16 15 14 13 12 11 05 04 23 10 06 03 24 21 08 25 02 07
01 26 27 28 29 30 31 22 09 20 19 35 34 33 32 36 18 17 16 15 14 13 12 11 05 04 23 10 06 03 24 21 08 25 02 07
01 26 27 28 29 30 31 32 33 34 35 19 20 09 22 36 18 17 16 15 14 13 12 11 05 04 23 10 06 03 24 21 08 25 02 07
01 26 27 28 29 30 31 32 33 34 35 36 22 09 20 19 18 17 16 15 14 13 12 11 05 04 23 10 06 03 24 21 08 25 02 07
01 26 27 28 29 30 31 32 33 34 35 36 22 09 24 03 06 10 23 04 05 11 12 13 14 15 16 17 18 19 20 21 08 25 02 07
01 26 27 28 29 30 31 32 33 34 35 36 22 09 08 21 20 19 18 17 16 15 14 13 12 11 05 04 23 10 06 03 24 25 02 07
01 26 27 28 29 30 31 32 33 34 35 36 08 09 22 21 20 19 18 17 16 15 14 13 12 11 05 04 23 10 06 03 24 25 02 07
01 26 27 28 29 30 31 32 33 34 35 36 08 09 22 21 20 19 18 17 16 15 14 13 12 11 05 04 03 06 10 23 24 25 02 07
01 26 27 28 29 30 31 32 33 34 35 36 08 09 22 21 20 19 18 17 16 15 14 13 12 11 05 04 03 02 25 24 23 10 06 07
01 02 03 04 05 11 12 13 14 15 16 17 18 19 20 21 22 09 08 36 35 34 33 32 31 30 29 28 27 26 25 24 23 10 06 07
01 02 03 04 05 11 12 13 14 15 16 17 18 19 20 21 22 09 08 07 06 10 23 24 25 26 27 28 29 30 31 32 33 34 35 36
01 02 03 04 05 06 07 08 09 22 21 20 19 18 17 16 15 14 13 12 11 10 23 24 25 26 27 28 29 30 31 32 33 34 35 36
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36

Die minimale erforderliche Anzahl ungerichteter Umordnungen zu bestimmen, ist algorithmisch nicht effizient machbar (sofern P ungleich NP ist – was aber fast jeder glaubt); auch nicht mit noch soviel Rechenpower; auch nicht in 1000 Jahren; auch nicht mit noch so raffinierten Algorithmen; auch nicht mit noch so toller Programmierkompetenz; auch nicht mit noch so tollen mathematischen Ideen:

***Das Problem ist NP-schwierig (technischer Begriff aus der Theoretischen Informatik),
und dies kann man auch beweisen.***

Mit zusätzlichen Vorzeichen versehen – beispielsweise mit lauter +-Vorzeichen – geht es aber wieder gut.



Man benötigt mindestens $36 + 1 - 3 + 0 + 0 = 34$ Umordnungen. Frage: Ist das kein Widerspruch zu den oben angegebenen 26 Umordnungen.

Thema 3: Sequenz-Alignment

Wie findet man Gene auf den Chromosomen? Suche in Datenbanken nach möglichst ähnlichen Sequenzen.

Zu welcher der beiden folgenden Sequenzen ist die Sequenz

ACCGTGCATACGTGCGTAATCCTCAATCGCG

ähnlicher?

TTCCGTCTTACGTTTCGTAATACTCACG

ACCTATGCATACGTGTTACGTACCTCAATTTCGG

Was heißt dabei „ähnlich“? Wie messen wir den Grad der Ähnlichkeit numerisch, sodass wir dann bezüglich einer solchen Maßzahl eine Maximierung vornehmen können?

- ✓ Einstreuen von Leerzeichen ist erlaubt: Mutationen können Zeichen löschen oder erzeugen.
- ✓ Bewerte zwei gleich lange Sequenzen mit eingestreuten Leerzeichen durch eine Scoring-Matrix:

Beispiel einer Scoring-Matrix:

	-	A	C	G	T
-		-2	-2	-2	-2
A	-2	1	-1	-1	-1
C	-2	-1	1	-1	-1
G	-2	-1	-1	1	-1
T	-2	-1	-1	-1	1

Beispiel einer realen Scoring-Matrix (BLOSUM-62) bei Proteinen (Sequenzen von Aminosäuren):

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-2
C	0	9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-2
D	-2	-3	6	2	-3	-1	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-2
E	-1	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-2
F	-2	-2	-3	-3	6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	3
G	0	-3	-1	-2	-3	6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-3
H	-2	-3	-3	0	-1	-2	8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	2
I	-1	-1	-1	-3	0	-4	-3	4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1
K	-1	-3	-3	1	-3	-2	-1	-3	5	-2	-1	0	-1	1	2	0	-1	-2	-3	-2
L	-1	-1	-1	-3	0	-4	-3	2	-2	4	2	-3	-3	-2	-2	-2	-1	1	-2	-1
M	-1	-1	-1	-2	0	-3	-2	1	-1	2	5	-2	-2	0	-1	-1	-1	1	-1	-1
N	-2	-3	-3	0	-3	0	1	-3	0	-3	-2	6	-2	0	0	1	0	-3	-4	-2
P	-1	-3	-3	-1	-4	-2	-2	-3	-1	-3	-2	-2	7	-1	-2	-1	-1	-2	-4	-3
Q	-1	-3	-3	2	-3	-2	0	-3	1	-2	0	0	-1	5	1	0	-1	-2	-2	-1
R	-1	-3	-3	0	-3	-2	0	-3	2	-2	-1	0	-2	1	5	-1	-1	-3	-3	-2
S	1	-1	-1	0	-2	0	-1	-2	0	-2	-1	1	-1	0	-1	4	1	-2	-3	-2
T	0	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	1	5	0	-2	-2
V	0	-1	-1	-2	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4	-3	-1
W	-3	-2	-2	-3	1	-2	-2	-3	-3	-2	-1	-4	-2	-2	-3	-3	-2	-3	11	2
Y	-2	-2	-2	-2	3	-3	2	-1	-2	-1	-1	-2	-1	-1	-2	-2	-2	-1	2	7

Definitionen

Sequenz $S = S(1)S(2)...S(i)...S(n)$

Sequenz $T = T(1)T(2).....T(j).....T(m)$

Ein ***Alignment*** von S mit T entsteht, indem man in S und T Leerzeichen – einstreut. Die entstehenden Sequenzen nennen wir S^* und T^* . Sie müssen gleich lang sein. Ferner dürfen nicht zwei Leerzeichen in S^* und T^* an derselben Position vorkommen:

$$S^* = S^*(1)S^*(2).....S^*(k)$$

$$T^* = T^*(1)T^*(2)....T^*(k)$$

Eine ***Scoringfunktion*** σ ordnet je zwei Zeichen a und b numerische Wert $\sigma(a,b)$, $\sigma(a,-)$ und $\sigma(-,b)$ zu.

In der Regel werden $\sigma(a,-)$, $\sigma(-,a)$ und $\sigma(a,b)$ mit $a \neq b$ negativ sein, während $\sigma(a,a)$ positiv ist.

Der *Wert eines Alignments* S^* und T^* der Länge k ist definiert als die Summe der $\sigma(S^*(i), T^*(i))$, für $i = 1, \dots, k$.

Gesucht ist ein Alignment S^* und T^* von S mit T , welches maximalen Wert hat.

Durchsuchen aller Alignments scheitert an der Suchraumgröße. Frage: *Wieviele Alignments gibt es?*

Intelligente Suche:

Sei $\sigma_{\max}(i, j)$ der maximale Wert, den ein Alignment der Anfangsstücke $S(1) \dots S(i)$ mit $T(1) \dots T(j)$ haben kann.

Eigentlich interessiert uns $\sigma_{\max}(n, m)$. Einen effizienten Algorithmus erhalten wir erstaunlicherweise aber dadurch, dass wir mehr berechnen, als wir eigentlich müssen, nämlich alle Werte $\sigma_{\max}(i, j)$, für $i = 1, \dots, n$ und $j = 1, \dots, m$. Außerdem berechnen wir noch $\sigma_{\max}(0, 0)$, $\sigma_{\max}(i, 0)$ und $\sigma_{\max}(0, j)$:

$$\sigma_{\max}(0,0) = 0$$

$$\sigma_{\max}(i,0) = \sigma(S(1),-) + \sigma(S(2),-) + \dots + \sigma(S(i),-)$$

$$\sigma_{\max}(0,j) = \sigma(-,T(1)) + \sigma(-,T(2)) + \dots + \sigma(-,T(j))$$

$\sigma_{\max}(i,j)$ = das Maximum der folgenden drei Zahlen:

$$\sigma(S(i),T(j)) + \sigma_{\max}(i-1,j-1)$$

$$\sigma(S(i),-) + \sigma_{\max}(i-1,j)$$

$$\sigma(-,T(j)) + \sigma_{\max}(i,j-1)$$

Die drei Zahlen spiegeln die drei Möglichkeiten wider, wie wir die rechten Sequenzen alignieren können:

S*(1)	S*(2)	...	S*(i-1)	S(i)
T*(1)	T*(2)	...	T*(j-1)	T(j)

oder

S*(1)	S*(2)	...	S*(i-1)	S(i)
T*(1)	T*(2)	...	T*(j)	-

oder

S*(1)	S*(2)	...	S*(i)	-
T*(1)	T*(2)	...	T*(j-1)	T(j)

Wir dürfen nun nicht den Fehler begehen, die obigen Formeln „rekursiv“ zu implementieren:

3-facher Unteraufruf $\Rightarrow$ Laufzeit 3^n

Stattdessen „iterativ“, am besten als Tabelle wie folgt, in der in jedem Feld auch noch die „Herkunft“ des Maximums protokolliert wird.

Beispiel mit S = ATCTGAT und T = TGCATA.

			T	G	C	A	T	A
0		0	← -2	← -4	← -6	← -8	← -10	← -12
1	A	↑ -2	← -1	← -3	← -5	← -5	← -7	← -9
2	T	↑ -4	← -1	← -2	← -4	← -6	← -4	← -6
3	C	↑ -6	↑ -3	← -2	← -1	← -3	← -5	← -5
4	T	↑ -8	↑ -5	← -4	← -3	← -2	← -2	← -4
5	G	↑ -10	↑ -7	← -4	← -5	← -4	← -3	← -3
6	A	↑ -12	↑ -9	← -6	← -5	← -4	← -5	← -2
7	T	↑ -14	↑ -11	← -8	← -7	← -6	← -3	← -4

Im Feld rechts unten lesen wir dann den Endwert -4 ab.

Ein Alignment mit maximalem Wert -4 kann durch Rückverfolgen der Pfeile bis zum Feld links oben konstruiert werden.

S* = T G C A T A -

T* = A T C T G A T

Weitere Themen hätten sein können :

Thema 4: Genom-Sequencing

Wie findet man die genaue Abfolge der Basen A, C, G, T auf den Chromosomen?

- Shotgun-Verfahren

Dabei ist das **Shortest Superstring Problem** zu lösen, welches – exemplarisch - wie folgt lautet:

In einem Satz der deutschen Sprache (alle Buchstaben klein geschrieben und ohne Zwischenräume zwischen den Wörtern) kommen folgende Fragmente vor (teilweise überlappend, eventuell mehrfach, nicht in der richtigen Reihenfolge):

nnendu derfo for denderin vialemetho annendu ichtt
rmatik kistsehrspa thodend tsehrsp bioi erinfo atik
denderin tikis ialeme bioinfor nderfordertni mat rdertnic
bioi annendun emetho form matikis nderfo ik rdertn
rivial nfor httri tsehrsp mati nforma chttriv

Man finde den kürzesten Satz, der all diese Fragmente enthält.
Die Hoffnung ist, dass dieser kürzeste Satz der „wahre“ Ori-
ginalsatz ist, aus dem die obigen Fragmente entnommen wur-
den.

Das Problem ist NP-schwierig (also nicht effizient exakt lös-
bar; wohl aber approximativ).

- DNA-Chips, Micro-Arrays

Weitere Details (auch zu den vorherigen Themen) können im
Vorlesungsskriptum „Bioinformatik“:

<http://www-lehre.inf.uos.de/bio01/>

nachgelesen werden.

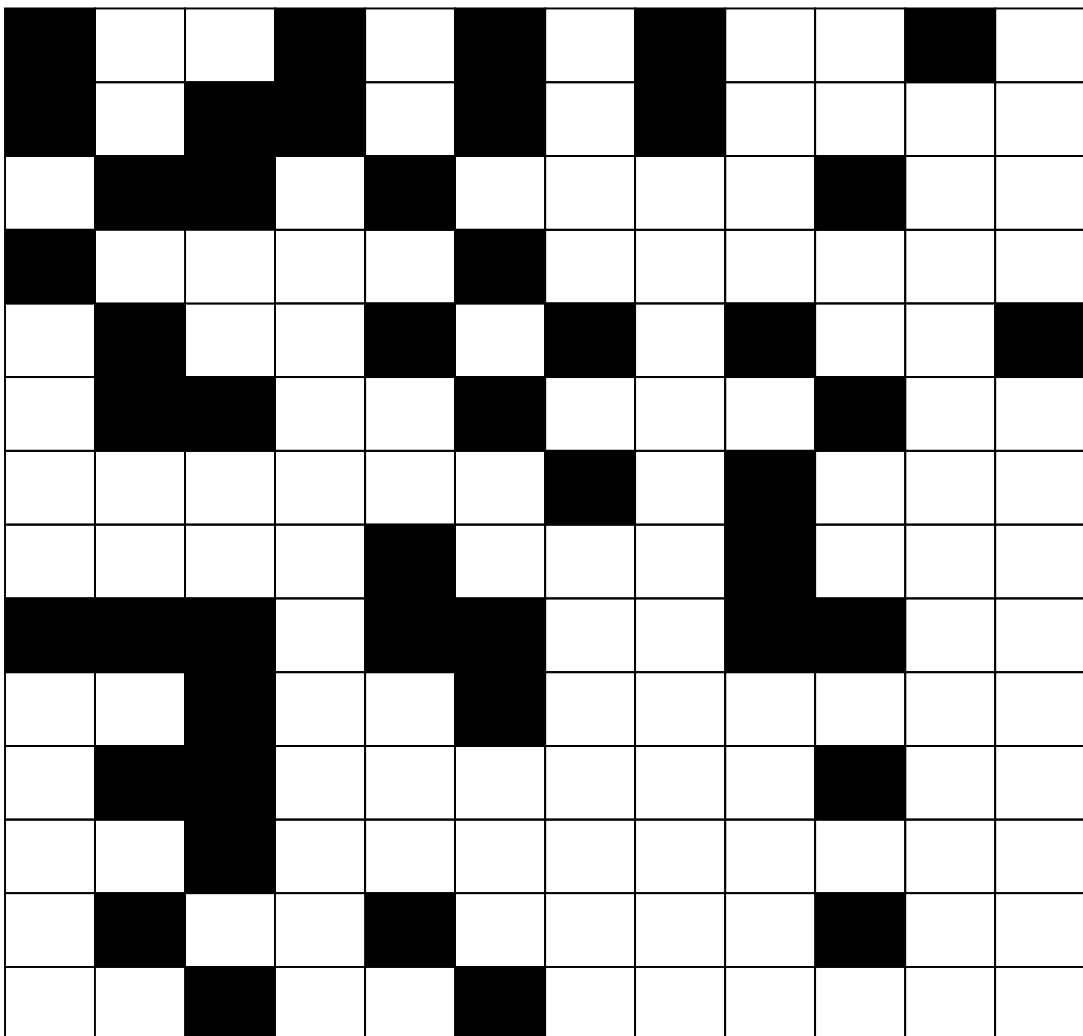
Thema 5: Genom-Mapping

Wie kann man zunächst eine grobe Karte eines Chromosoms erstellen?

- DOUBLE-DIGEST-Verfahren: NP-schwierig
- SINGLE-DIGEST-Verfahren: im Mittel effizient
- MAPPING-BY-HYBRIDIZATION: effizient lösbar

CONSECUTIVE ONES Problem

Man ordne die Spalten einer binären Matrix so um, dass am Ende in jeder Zeile alle Einsen nebeneinander stehen.



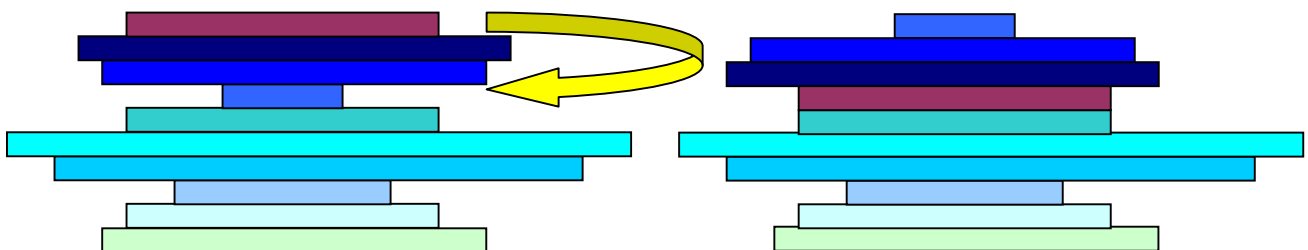
Zum Abschluss noch etwas zur Entspannung:

Ein Spezialfall des Genom-Rearrangement-Problems ist das

PANCACE FLIPPING Problem:

Stapel von n unterschiedlich großen Pfannkuchen, den wir auf einer Hand halten.

Greife mit der anderen Hand beliebig in den Stapel und kippe den herausgegriffenen Teil in umgekehrter Reihenfolge auf den Reststapel.



Mit wie vielen solcher Operationen kann man n Pfannkuchen der Größe nach ordnen? Obere Schranken – Untere Schranken

B. Gates (1979,...)