Approximative Gencluster

(kaum: Genome Rearrangement)

– Aktuelle Entwicklungen
aus Bioinformatik und komparativer Genomforschung –

Jens Stoye

AG Genominformatik, Technische Fakultät

Institut für Bioinformatik, Centrum für Biotechnologie (CeBiTec)

Universität Bielefeld

Überblick

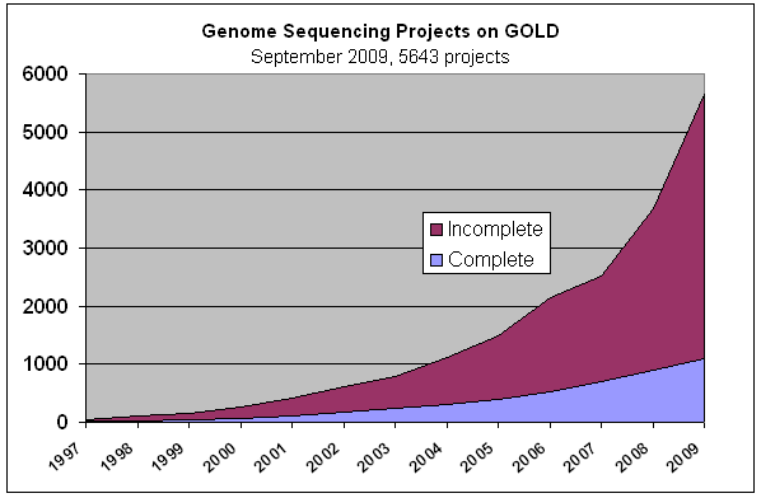
1. Komparative Genomanalyse
2. Approximative Gencluster
3. Experimentelle Ergebnisse
4. Rekonstruktion phylogenetischer Bäume
5. Zusammenfassung und Ausblick

Überblick

1. Komparative Genomanalyse
2. Approximative Gencluster
3. Experimentelle Ergebnisse
4. Rekonstruktion phylogenetischer Bäume
5. Zusammenfassung und Ausblick

Komparative Genomik

- Vergleich des genetischen Materials verschiedener Organismen
- Konservierungsmuster durch selektive Evolution
- Klassischer Ansatz: Vergleich einzelner Gene
- Neuere Entwicklungen: Vergleich kompletter Genome
- Idee: Zusammenhang von Genom-Struktur und Genfunktion



(Quelle: `http://genomesonline.org/gold_statistics.htm`)

Genom-Modell

genome 1: ...cgtaggtctacgcctaggcttcagtcgtattgatactcgtagttgcttacgtagcatgatcagtctgctgagtcgtacg...

genome 2: ...cgtacagtctacgtcaacgttcacgtattgatgccctcgtagtcacgctacgtacgtaatgctgagactcatcgtacg...

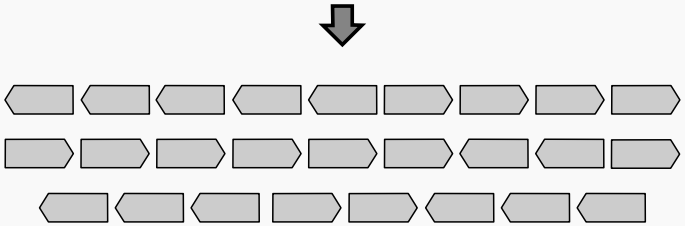
genome 3: ...cgtagatgagtctaagtcgtattgatactcggttgagtacgtaggtacatgatgtgctaagagactgtcgtcgtacg...

Genom-Modell

genome 1: ...cgtaggtctacgcctaggcttcagtcgtattgatactcgtagttgcttacgtagcatgatcagtctgctgagtcgtacg...

genome 2: ...cgtacagtctacgtcaacgttcacgtattgatgccctcgtagtcacgctacgtacgtaatgctgagactcatcgtacg...

genome 3: ...cgtagatgagtctaagtcgtattgatactcggttgagtacgtaggtacatgatgtgctaagagactgtcgtcgtacg...



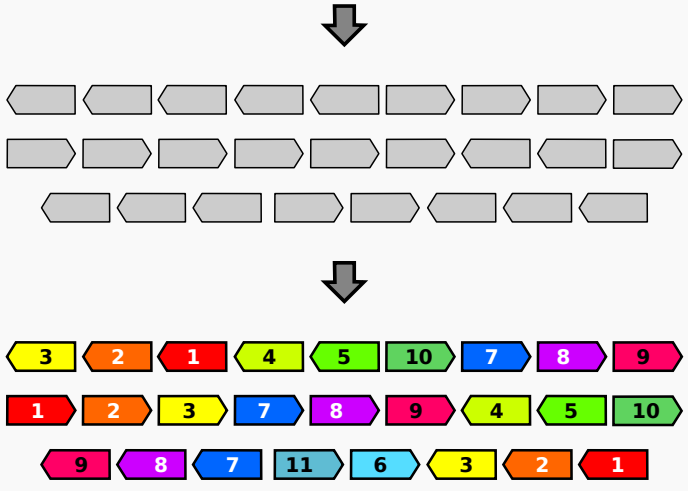
sequence of genes

Genom-Modell

genome 1: ...cgtaggtctacgcctaggcttcagtcgtattgatactcgtagttgcttacgtagcatgatcagtctgctgagtcgtacg...

genome 2: ...cgtacagtctacgtcaacgttcacgtattgatgccctcgtagtcacgctacgtacgtaatgctgagactcatcgtacg...

genome 3: ...cgtagatgagtctaagtcgtattgatactcggttgagtacgtaggtacatgatgtgctaagagactgtcgtcgtacg...



sequence of genes

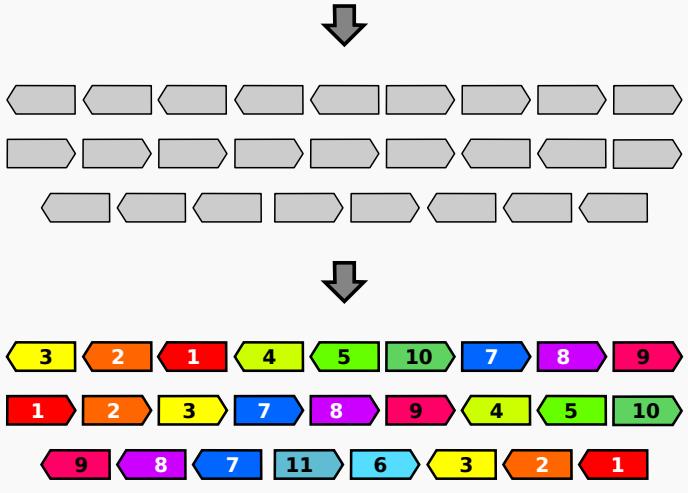
homology based labeling of genes

Genom-Modell

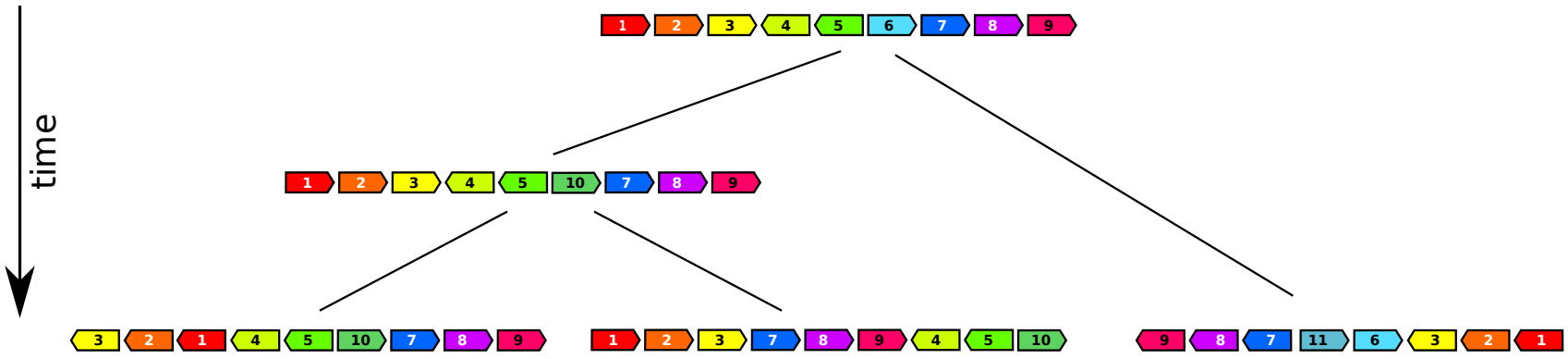
genome 1: ...cgtaggtctacgcctaggcttcagtcgtattgatactcgtagttgcttacgtagcatgatcagtctgctgagtcgtacg...

genome 2: ...cgtacagtctacgtcaacgttcacgtattgatgccctcgtagtcacgctacgtacgtaatgctgagactcatcgtacg...

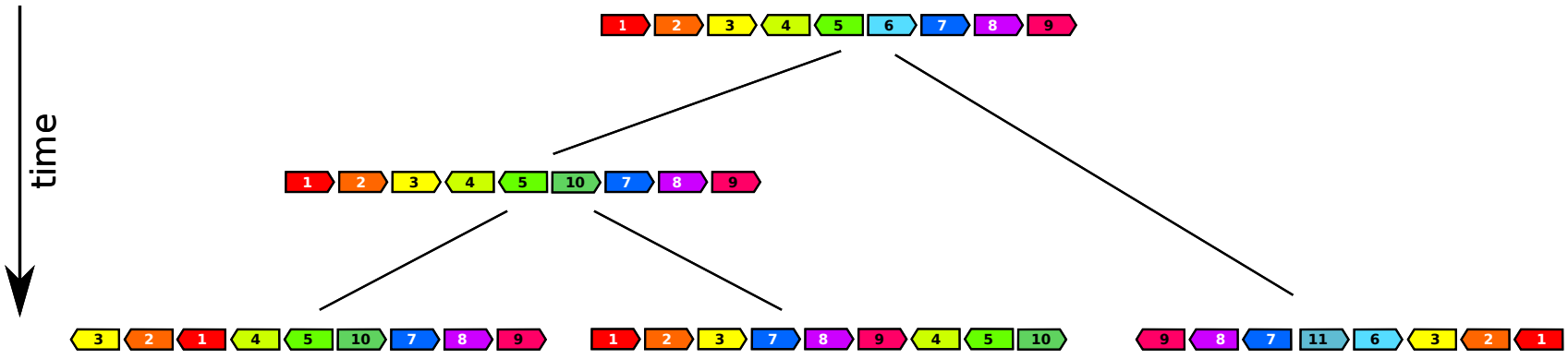
genome 3: ...cgtagatgagtctaagtcgtattgatactcggttgagtacgtaggtacatgatgtgctaagagactgtcgtcgtacg...



Evolution kompletter Genome



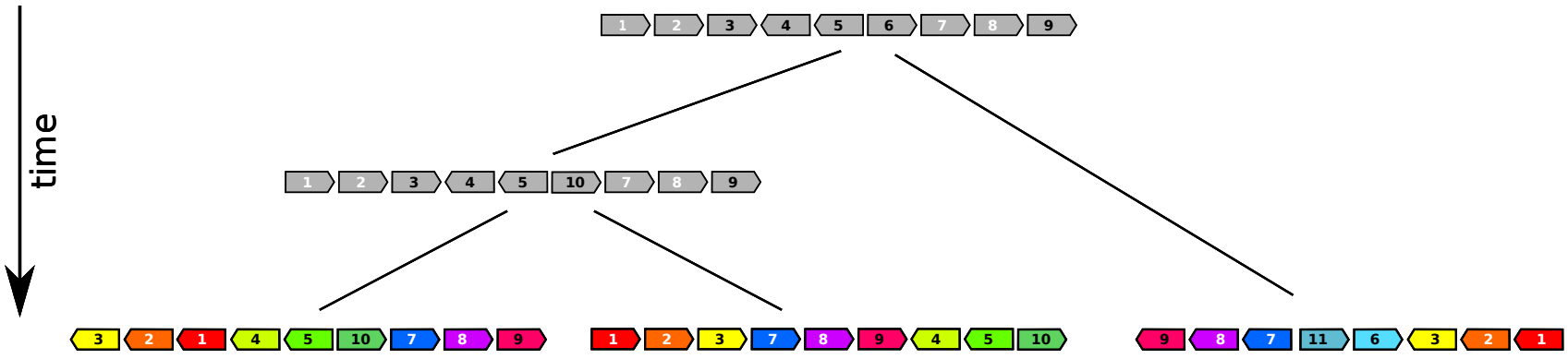
Evolution kompletter Genome



Zwei komplementäre Kräfte beeinflussen die Genreihenfolge:

- zufällige Umordnungen $\longrightarrow$ phylogenetische Beziehungen
- Selektionsdruck $\longrightarrow$ funktionelle Beziehungen

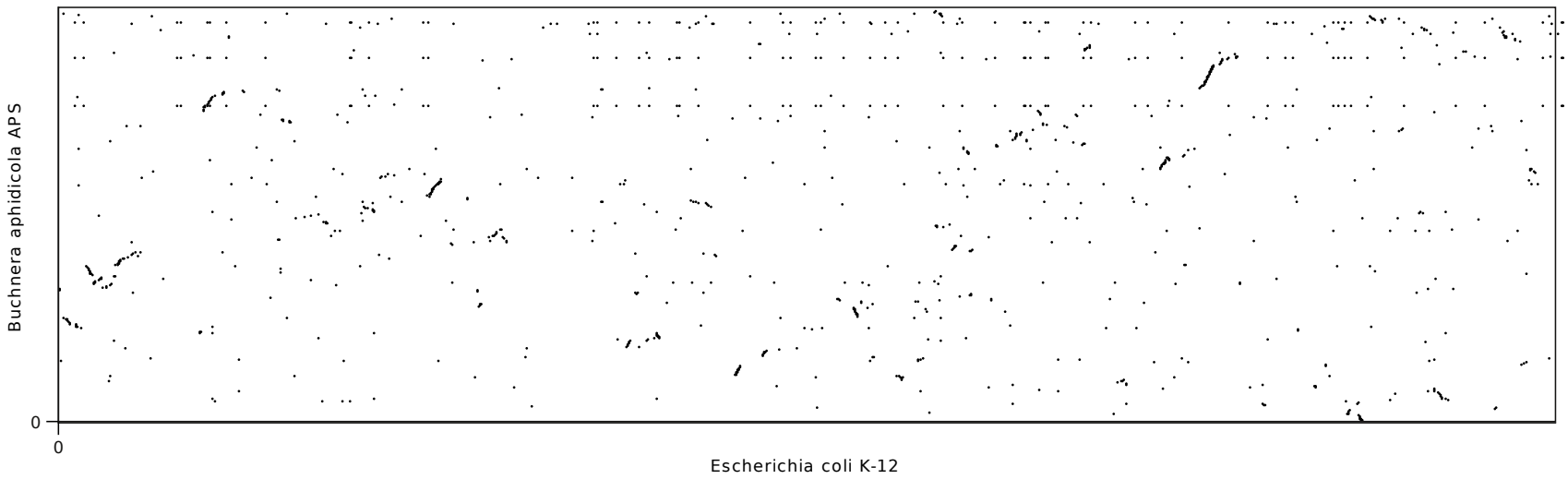
Evolution kompletter Genome



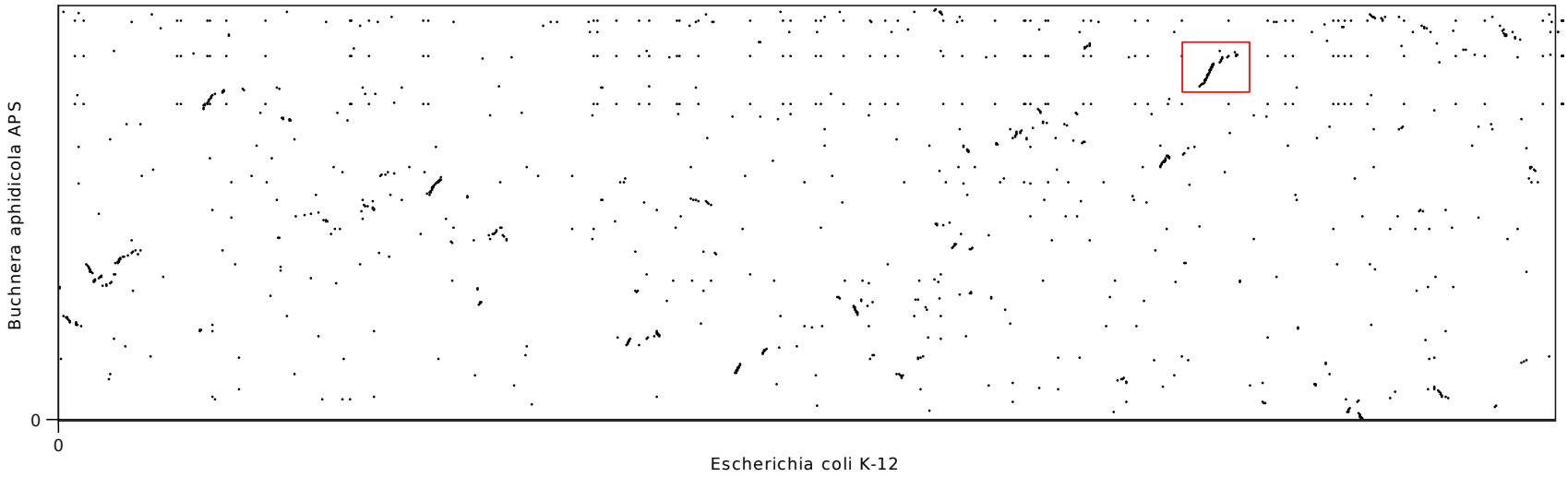
Zwei komplementäre Kräfte beeinflussen die Genreihenfolge:

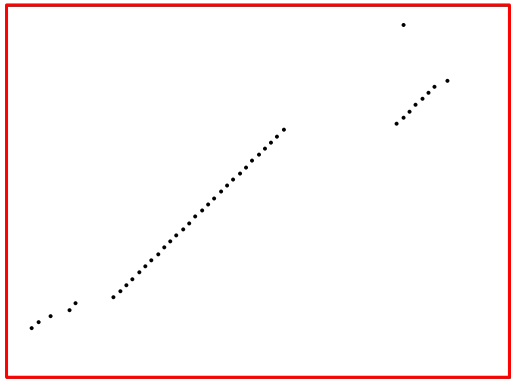
- zufällige Umordnungen $\longrightarrow$ phylogenetische Beziehungen
- Selektionsdruck $\longrightarrow$ funktionelle Beziehungen

Dot-Plot der Genreihenfolge

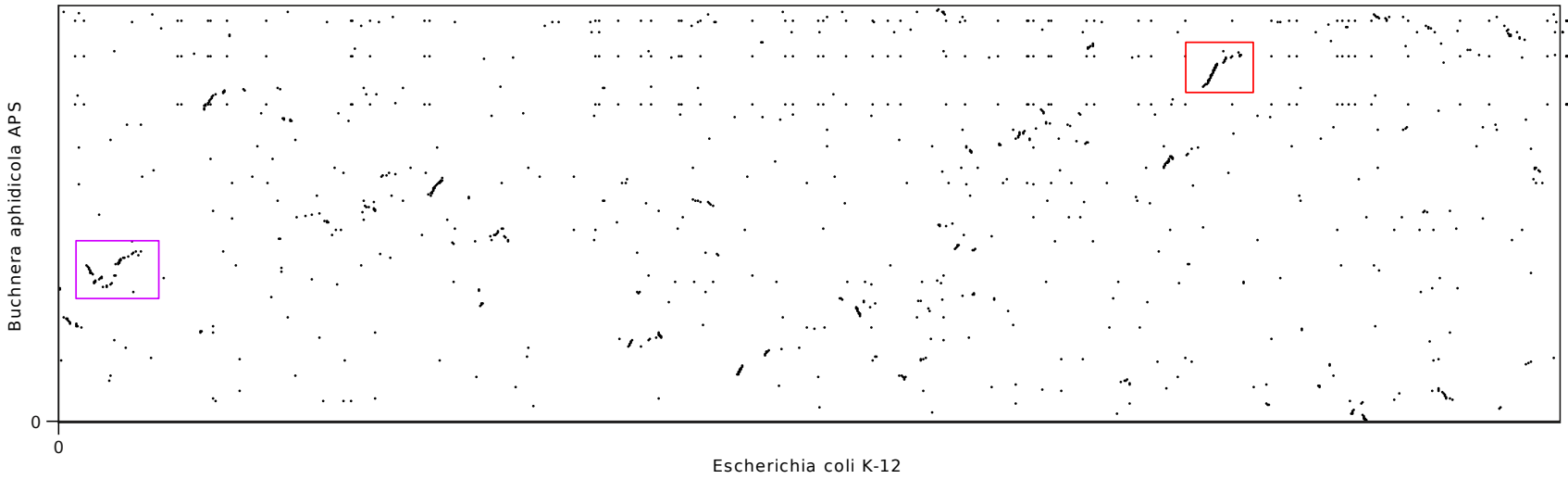


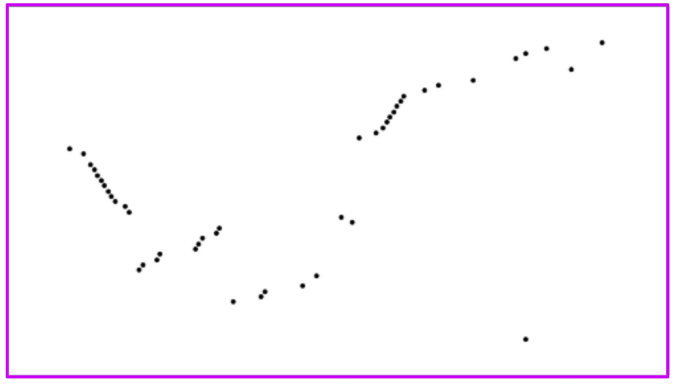
Dot-Plot der Genreihenfolge

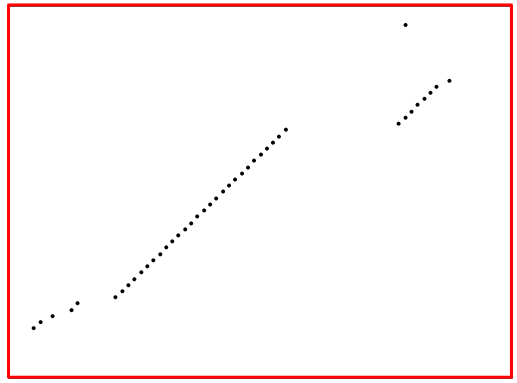




Dot-Plot der Genreihenfolge



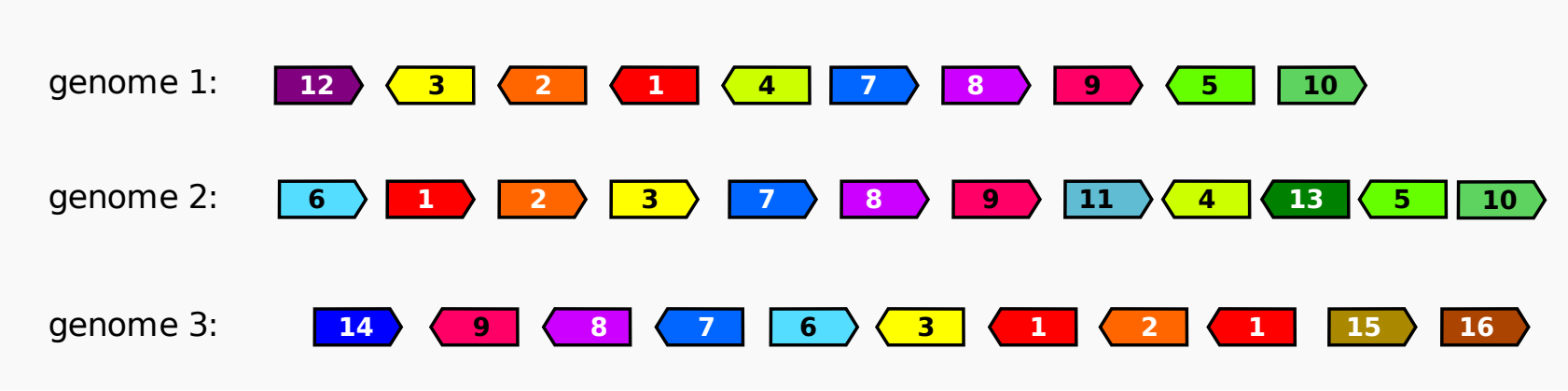




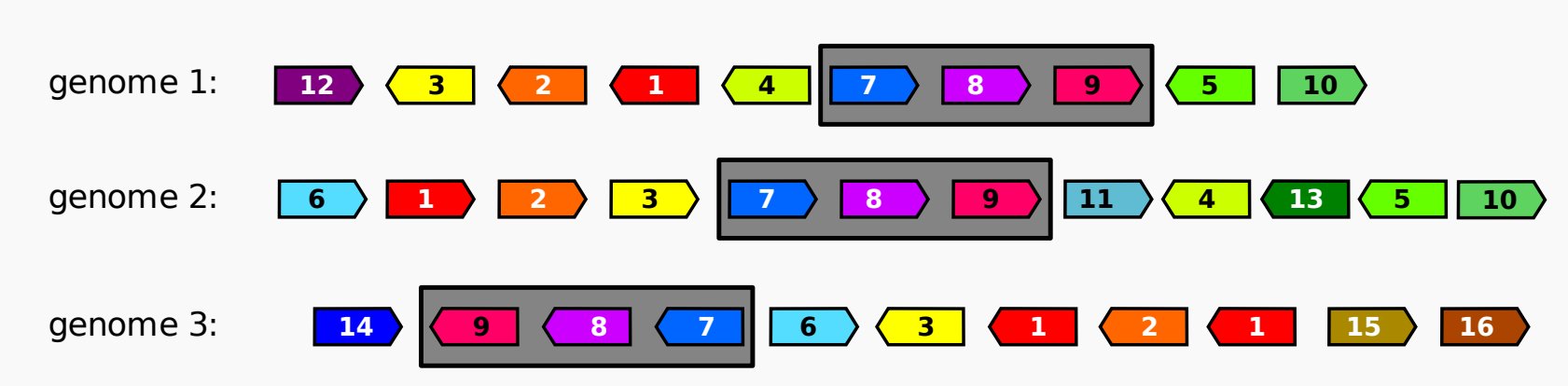
Überblick

1. Komparative Genomanalyse
2. Approximative Gencluster
   - Referenz-Gencluster
   - Median-Gencluster
   - Center-Gencluster
3. Experimentelle Ergebnisse
4. Rekonstruktion phylogenetischer Bäume
5. Zusammenfassung und Ausblick

Gencluster

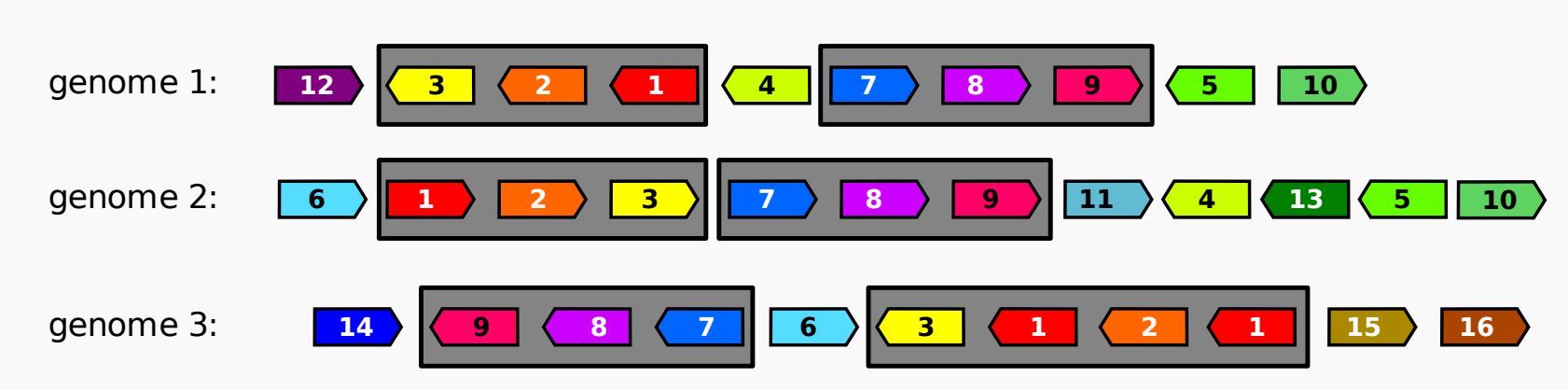

genome 1:
12
3
2
1
4
7
8
9
5
10
genome 2:
6
1
2
3
7
8
9
11
4
13
5
10
genome 3:
14
9
8
7
6
3
1
2
1
15
16

Gencluster



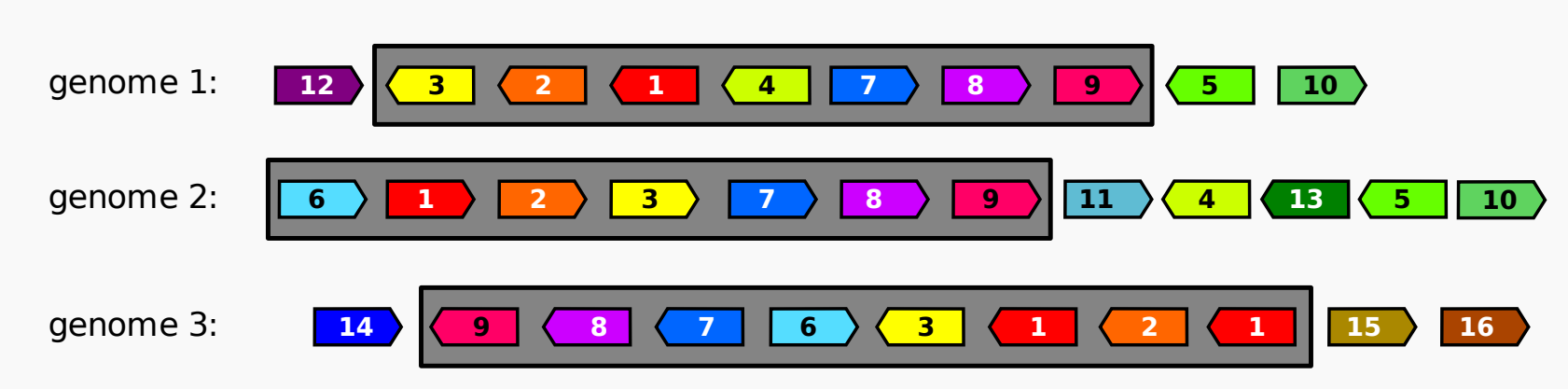
- Kollineare Gencluster:
  - Menge der Gene liegen *en bloc* in mehreren Genomen.
  - Gen-Reihenfolge ist komplett erhalten (abgesehen von Inversionen).

Gencluster



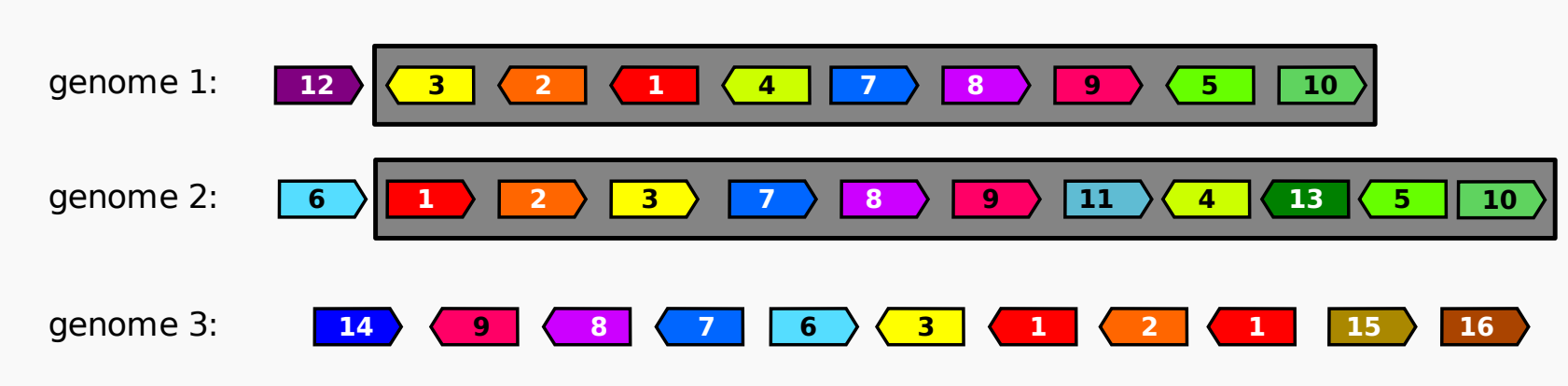
- Kollineare Gencluster:
  - Menge der Gene liegen *en bloc* in mehreren Genomen.
  - Gen-Reihenfolge ist komplett erhalten (abgesehen von Inversionen).
- Perfekte Gencluster (*common intervals*):
  - Genreihenfolge innerhalb des Block wird vernachlässigt.
  - Multiple Vorkommen eines Gens im Cluster sind möglich.

Gencluster



- Approximative Gencluster:
  - Zusätzliche und fehlende Gene sind möglich.

Gencluster



- Approximative Gencluster:
  - Zusätzliche und fehlende Gene sind möglich.
- q-abdeckende Gencluster:
  - Das Gencluster muss nur in einem Teil der Genome vorkommen.

Modelle für approximative Gencluster

- *r*-Fenster (Friedman and Hughes, 2001)
  - feste Block-Größe
  - Schnittmenge der Gene der einzelnen Vorkommen
  - exponentielle Laufzeit bezüglich der Anzahl Genome
- max-Lücke (Bergeron, Corteel and Raffinot, 2002)
  - obere Schranke für Insertionslänge
  - Schnittmenge der Gene der einzelnen Vorkommen
  - exponentielle Laufzeit bezüglich der Anzahl Genome
- Approximative Gencluster (Rahmann and Klau, 2006)
  - sehr allgemeines Model, schließt die meisten anderen mit ein
  - *Integer Linear Programming*-Lösung existiert (exponentiell)
- Referenz-Gencluster
- Median-Gencluster
- Center-Gencluster

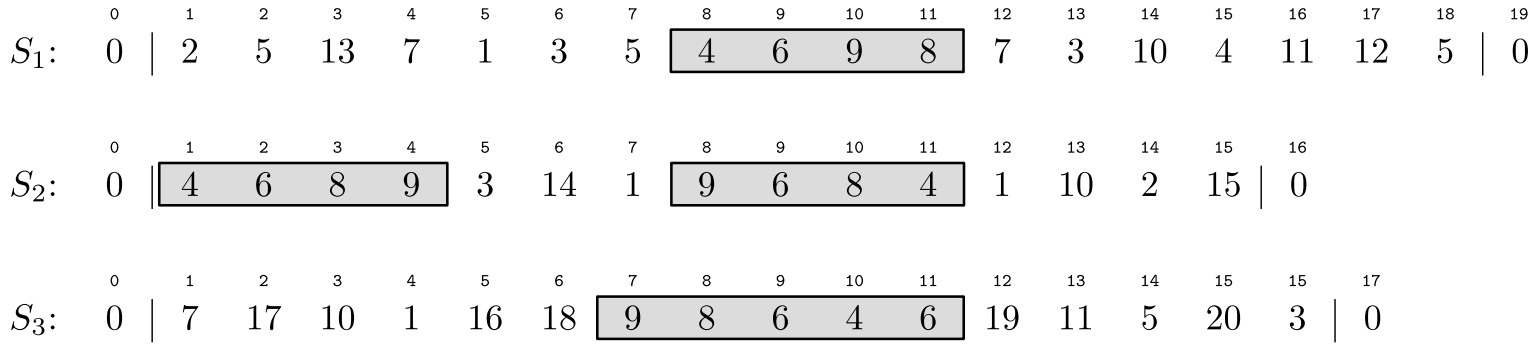
Grundlegende Definitionen

- Genome: Menge von k Sequenzen $S_1, \ldots, S_k$ über Alphabet Σ
- $n =$ maximale Länge der $S_1, \ldots, S_k$, $|\Sigma| \in \Theta(n)$
- **Zeichenmenge** eines Teilwortes:

$$\mathcal{CS}(S_\ell[i_\ell, j_\ell]) = \{S_\ell[i] \mid i_\ell \leq i \leq j_\ell\}$$

- **Gemeinsame Intervalle**:

$$([i_1, j_1], \ldots, [i_k, j_k]) \text{ mit } \mathcal{CS}(S_1[i_1, j_1]) = \ldots = \mathcal{CS}(S_k[i_k, j_k])$$



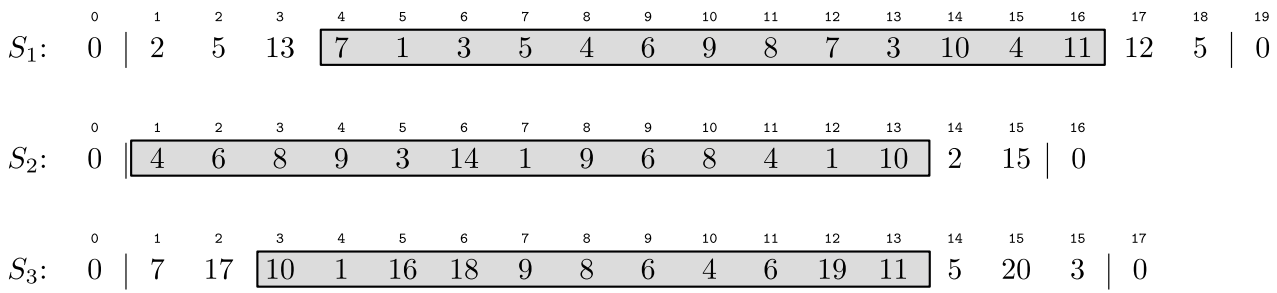
Grundlegende Definitionen

- **Symmetrische Mengen-Distanz**:

$$D(C, C') = |C \setminus C'| + |C' \setminus C|$$

- δ-**Ort** von $C \subseteq \Sigma$: $[i_\ell, j_\ell]_{S_\ell}$ mit $D(C, \mathcal{CS}(S_\ell[i_\ell, j_\ell])) \leq \delta$
- **Approximative gemeinsame Intervalle**: $([i_1, j_1], \ldots, [i_k, j_k])$ mit einer Menge $C \subseteq \Sigma$ für einen **Schwellwert** δ:

$$\sum_{\ell=1}^{k} D\big(C, \mathcal{CS}(S_\ell[i_\ell, j_\ell])\big) \leq \delta$$



Überblick

1. Komparative Genomanalyse
2. Approximative Gencluster
   - Referenz-Gencluster
   - Median-Gencluster
   - Center-Gencluster
3. Experimentelle Ergebnisse
4. Rekonstruktion phylogenetischer Bäume
5. Zusammenfassung und Ausblick

Variante 1: Referenz-Gencluster — Problemdefinition

Gegeben:

- Sequenzen $S_1, \ldots, S_k$ über dem Alphabet der Gene Σ
- s (minimale Clustergröße)
- δ (Distanz-Schwellwert)

Gesucht: alle $M \subseteq \Sigma$ mit

- $M = \mathcal{CS}(S_m[i_m, j_m])$ für ein $S_m[i_m, j_m]$, $1 \leq m \leq k$
- $\sum_{\ell=1}^{k} D\big(M, \mathcal{CS}(S[i_\ell, j_\ell])\big) \leq \delta$ für eine Kombination von Teilworten $(S_1[i_1, j_1], \ldots, S_k[i_k, j_k])$
- $|M| \geq s$

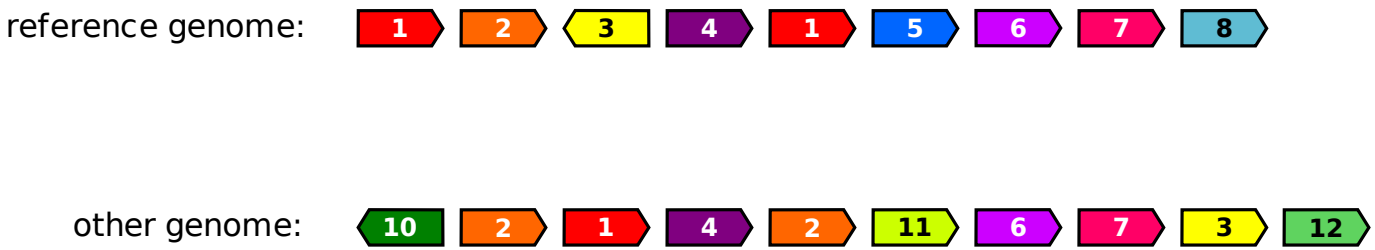
Eine solche Menge M heißt Referenz-Gencluster von $S_1, \ldots, S_k$.

Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
  $\longrightarrow$ Startpunkt der Berechnung
- Naiver Ansatz: berechne für alle Teilworte die paarweisen Distanzen zu allen anderen Teilworten $\mathcal{O}(k^2 n^4)$
- Unser Ansatz: Erweiterung des Algorithmus *Connecting Intervals* (Schmidt & JS, 2004)

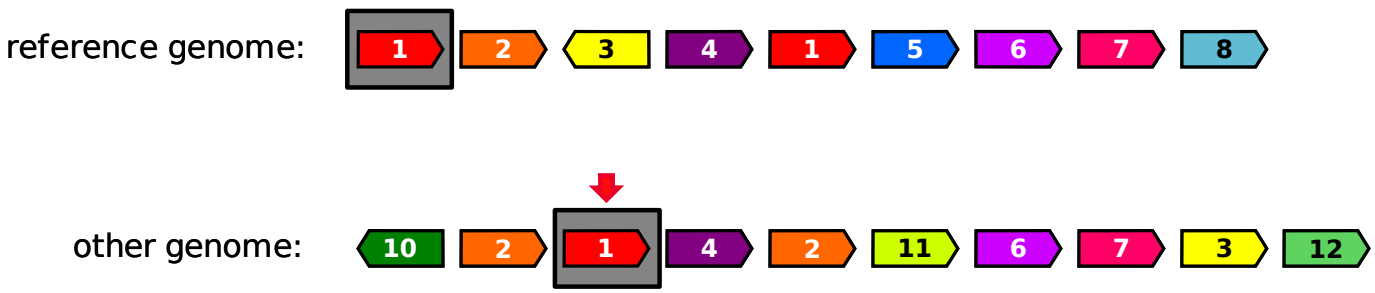
Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
  $\rightarrow$ Startpunkt der Berechnung
- Naiver Ansatz: berechne für alle Teilworte die paarweisen Distanzen zu allen anderen Teilworten $\mathcal{O}(k^2 n^4)$
- Unser Ansatz: Erweiterung des Algorithmus *Connecting Intervals* (Schmidt & JS, 2004)
- Beispiel: $s = 6$, $\delta = 2$



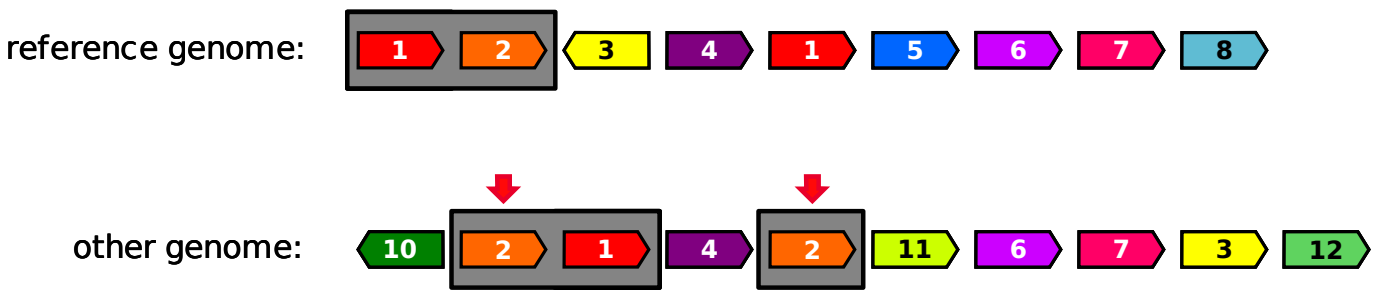
Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
  $\rightarrow$ Startpunkt der Berechnung
- Naiver Ansatz: berechne für alle Teilworte die paarweisen Distanzen zu allen anderen Teilworten $\mathcal{O}(k^2 n^4)$
- Unser Ansatz: Erweiterung des Algorithmus *Connecting Intervals* (Schmidt & JS, 2004)
- Beispiel: $s = 6$, $\delta = 2$



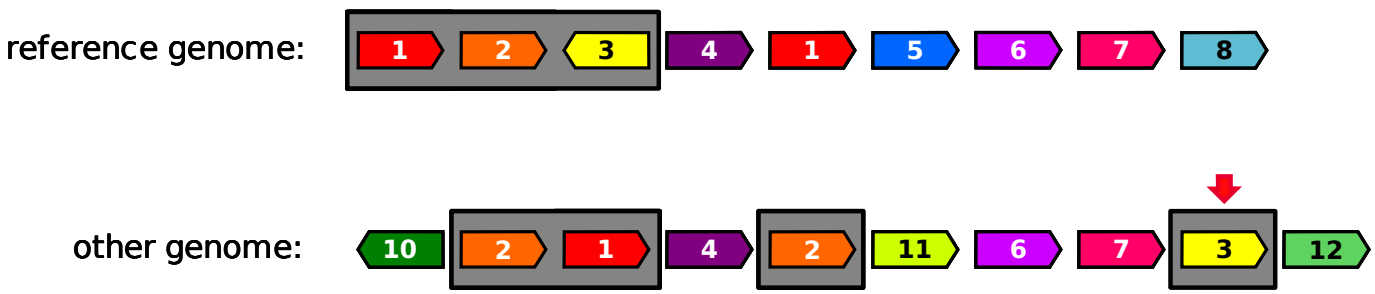
Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
  $\rightarrow$ Startpunkt der Berechnung
- Naiver Ansatz: berechne für alle Teilworte die paarweisen Distanzen zu allen anderen Teilworten $\mathcal{O}(k^2 n^4)$
- Unser Ansatz: Erweiterung des Algorithmus *Connecting Intervals* (Schmidt & JS, 2004)
- Beispiel: $s = 6$, $\delta = 2$



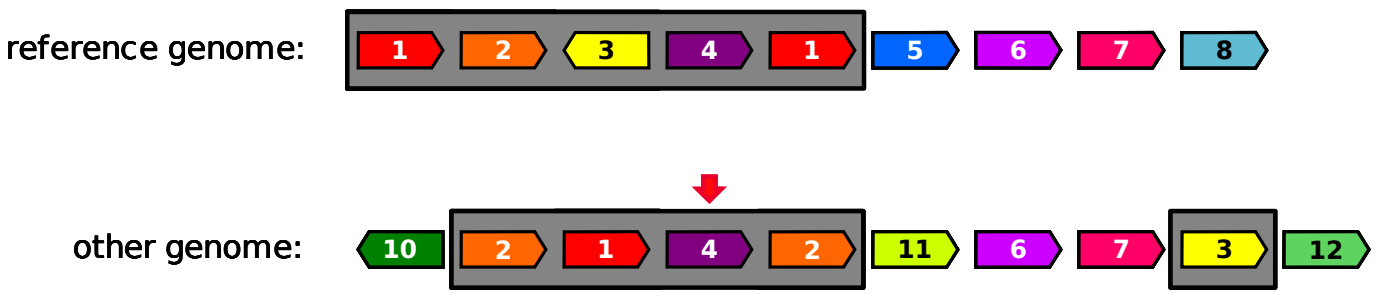
Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
  $\rightarrow$ Startpunkt der Berechnung
- Naiver Ansatz: berechne für alle Teilworte die paarweisen Distanzen zu allen anderen Teilworten $\mathcal{O}(k^2 n^4)$
- Unser Ansatz: Erweiterung des Algorithmus *Connecting Intervals* (Schmidt & JS, 2004)
- Beispiel: $s = 6$, $\delta = 2$



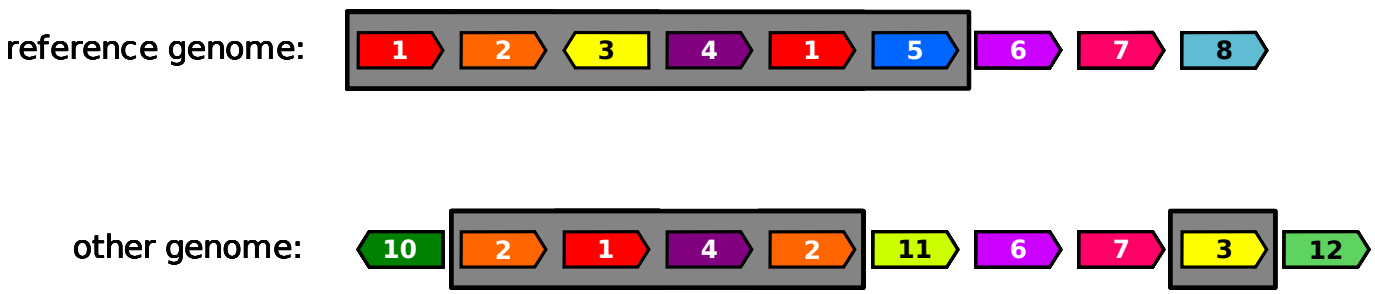
Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
  $\rightarrow$ Startpunkt der Berechnung
- Naiver Ansatz: berechne für alle Teilworte die paarweisen Distanzen zu allen anderen Teilworten $\mathcal{O}(k^2 n^4)$
- Unser Ansatz: Erweiterung des Algorithmus *Connecting Intervals* (Schmidt & JS, 2004)
- Beispiel: $s = 6$, $\delta = 2$



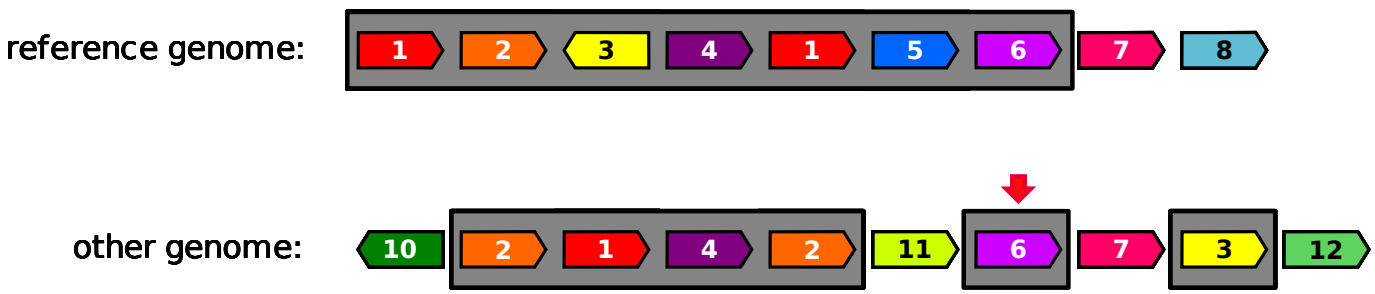
Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
  $\rightarrow$ Startpunkt der Berechnung
- Naiver Ansatz: berechne für alle Teilworte die paarweisen Distanzen zu allen anderen Teilworten $\mathcal{O}(k^2 n^4)$
- Unser Ansatz: Erweiterung des Algorithmus *Connecting Intervals* (Schmidt & JS, 2004)
- Beispiel: $s = 6$, $\delta = 2$



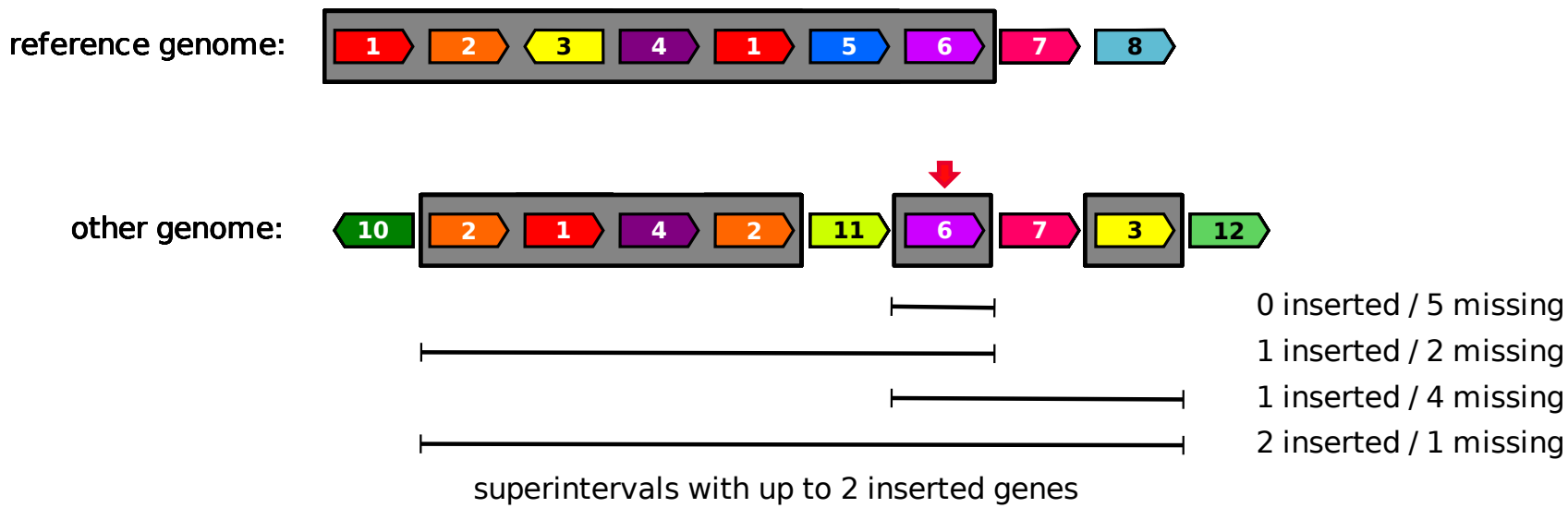
Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
  $\rightarrow$ Startpunkt der Berechnung
- Naiver Ansatz: berechne für alle Teilworte die paarweisen Distanzen zu allen anderen Teilworten $\mathcal{O}(k^2 n^4)$
- Unser Ansatz: Erweiterung des Algorithmus *Connecting Intervals* (Schmidt & JS, 2004)
- Beispiel: $s = 6$, $\delta = 2$



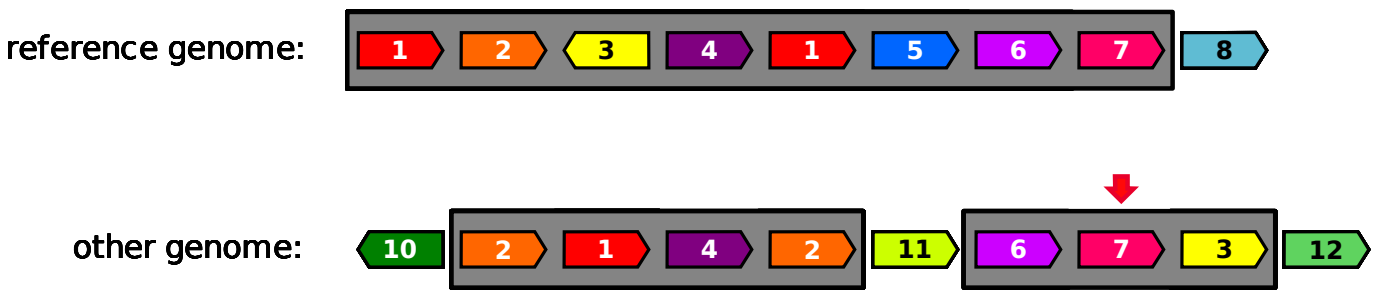
Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
  $\rightarrow$ Startpunkt der Berechnung
- Naiver Ansatz: berechne für alle Teilworte die paarweisen Distanzen zu allen anderen Teilworten $\mathcal{O}(k^2 n^4)$
- Unser Ansatz: Erweiterung des Algorithmus *Connecting Intervals* (Schmidt & JS, 2004)
- Beispiel: $s = 6$, $\delta = 2$



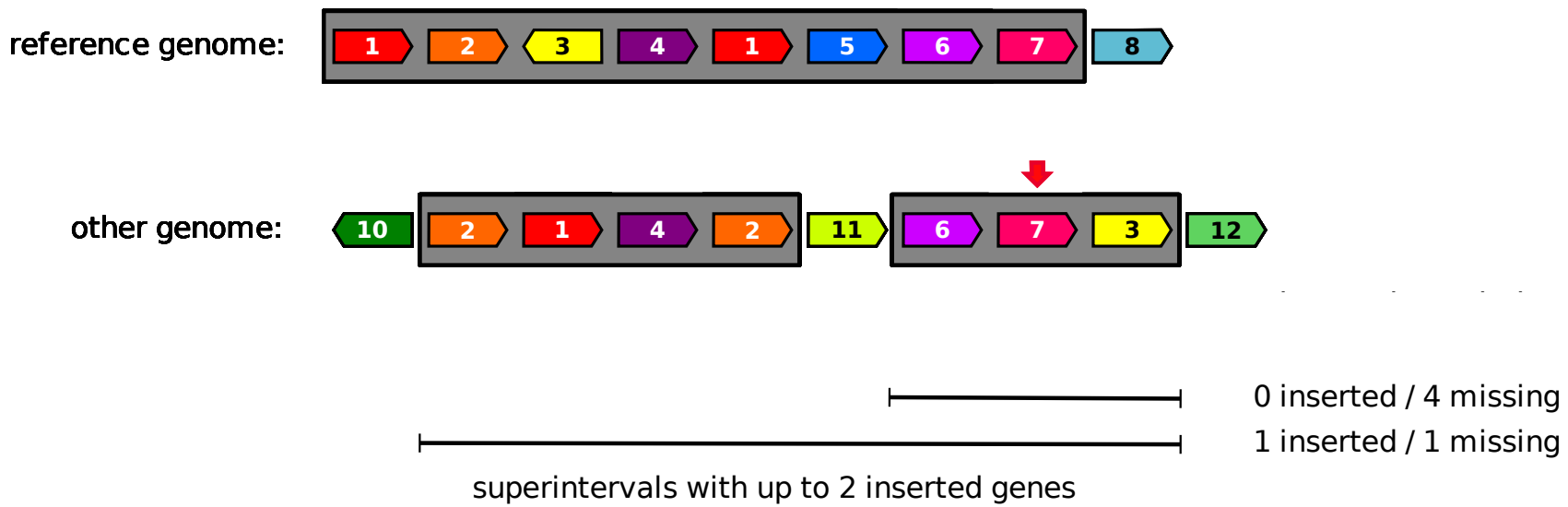
Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
  $\rightarrow$ Startpunkt der Berechnung
- Naiver Ansatz: berechne für alle Teilworte die paarweisen Distanzen zu allen anderen Teilworten $\mathcal{O}(k^2 n^4)$
- Unser Ansatz: Erweiterung des Algorithmus *Connecting Intervals* (Schmidt & JS, 2004)
- Beispiel: $s = 6$, $\delta = 2$



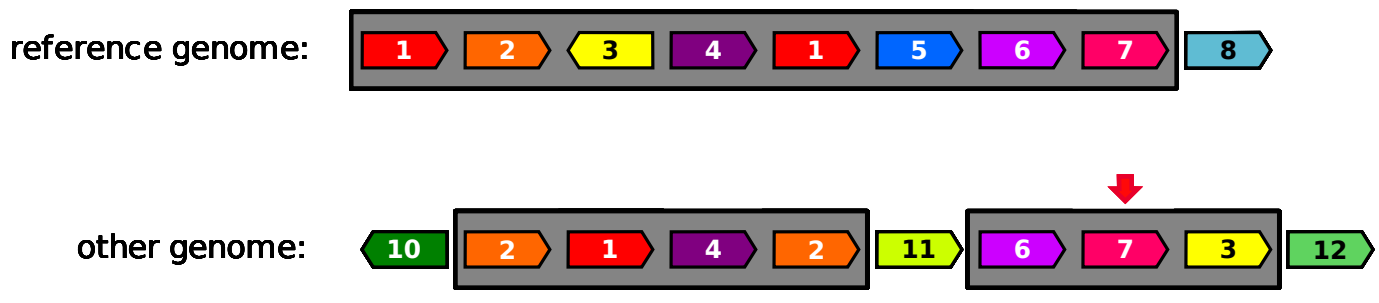
Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
  $\rightarrow$ Startpunkt der Berechnung
- Naiver Ansatz: berechne für alle Teilworte die paarweisen Distanzen zu allen anderen Teilworten $\mathcal{O}(k^2 n^4)$
- Unser Ansatz: Erweiterung des Algorithmus *Connecting Intervals* (Schmidt & JS, 2004)
- Beispiel: $s = 6$, $\delta = 2$



Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
  $\rightarrow$ Startpunkt der Berechnung
- Naiver Ansatz: berechne für alle Teilworte die paarweisen Distanzen zu allen anderen Teilworten $\mathcal{O}(k^2 n^4)$
- Unser Ansatz: Erweiterung des Algorithmus *Connecting Intervals* (Schmidt & JS, 2004)
- Beispiel: $s = 6$, $\delta = 2$



- Laufzeit: $\mathcal{O}(k^2 n^2 (1 + \delta^2))$, Platz: $\mathcal{O}(k n^2)$

Überblick

1. Komparative Genomanalyse
2. Approximative Gencluster
   - Referenz-Gencluster
   - Median-Gencluster
   - Center-Gencluster
3. Experimentelle Ergebnisse
4. Rekonstruktion phylogenetischer Bäume
5. Zusammenfassung und Ausblick

Variante 2: Median-Gencluster

- Referenz-Gencluster können in polynomieller Zeit berechnet werden
- Wieso noch Alternativen betrachten?
  $\longrightarrow$ Kein Optimalitätskriterium für den Konsensus
- Das perfekte Vorkommen ist eine nahe verwandte Referenzmenge, aber nicht notwendigerweise die am nächsten verwandte.
- Idee: Kombinationen von approximativen Vorkommen
  $\longrightarrow$ Finde eine Konsensus-Menge, die die Abstände zu allen Vorkommen minimiert.
- Kein perfektes Vorkommen der Konsensus-Menge!
- Exponentieller Suchraum: $\mathcal{O}(n^{2k})$ Teilwort-Kombinationen, $\mathcal{O}(2^{|\Sigma|})$ mögliche Konsensus-Mengen

Variante 2: Median-Gencluster — Problemdefinition

Gegeben:

- Sequenzen $S_1, \ldots, S_k$ über dem Alphabet der Gene Σ
- s (minimale Clustergröße)
- δ (Distanz-Schwellwert)

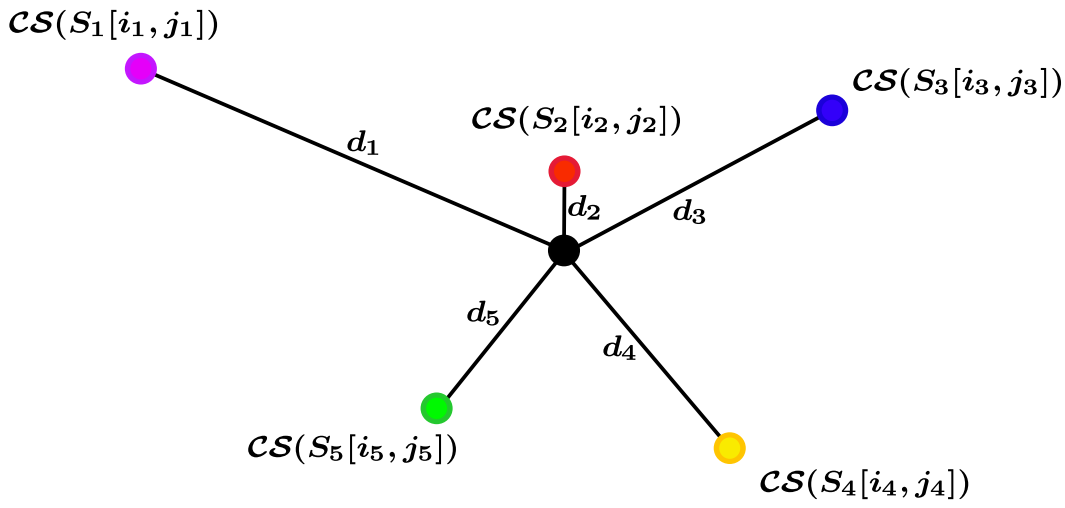
Gesucht: alle $M \subseteq \Sigma$ mit

- M ist Median für ein $\mathcal{CS}(S_1[i_1, j_1]), ..., \mathcal{CS}(S_k[i_k, j_k])$
- $\sum_{\ell=1}^{k} D\big(M, \mathcal{CS}(S[i_\ell, j_\ell])\big) \leq \delta$
- $|M| \geq s$

Eine solche Menge M heißt Median-Gencluster von $S_1, \ldots, S_k$.

Variante 2: Median-Gencluster — Reduktion des Suchraums

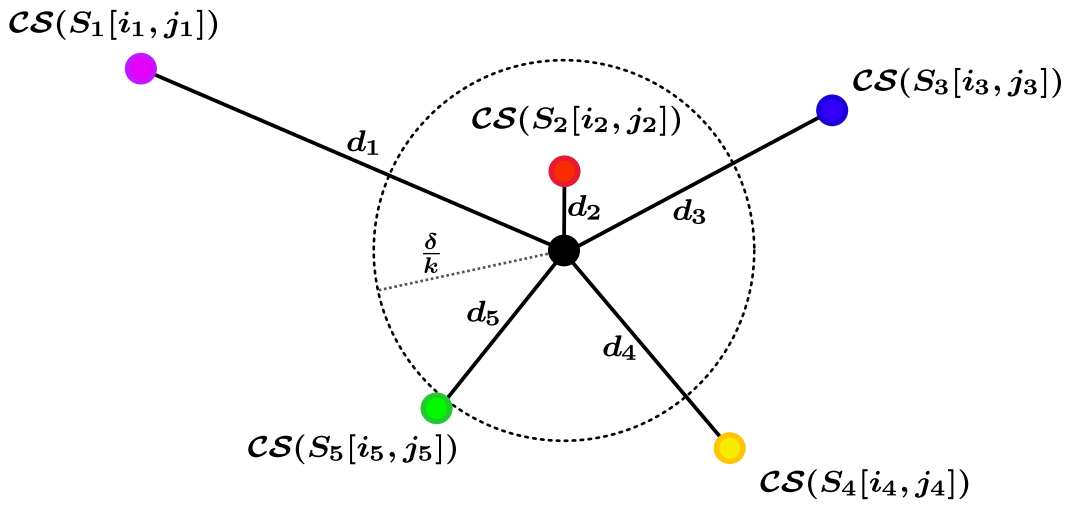
- Suchraum: $\mathcal{O}(n^{2k})$ Kombinationen von Teilworten von $S_1, ..., S_k$
- Cluster-Filter:



- Median-Distanzschwellwert: $\sum_{\ell=1}^{k} d_\ell \leq \delta$

Variante 2: Median-Gencluster — Reduktion des Suchraums

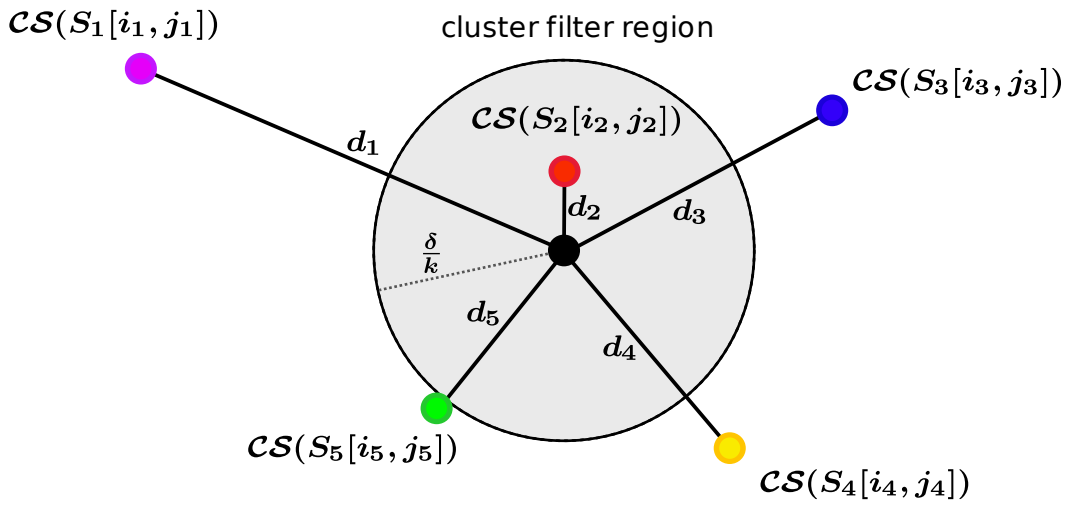
- Suchraum: $\mathcal{O}(n^{2k})$ Kombinationen von Teilworten von $S_1, ..., S_k$
- Cluster-Filter:



- Median-Distanzschwellwert: $\sum_{\ell=1}^{k} d_\ell \leq \delta$

Variante 2: Median-Gencluster — Reduktion des Suchraums

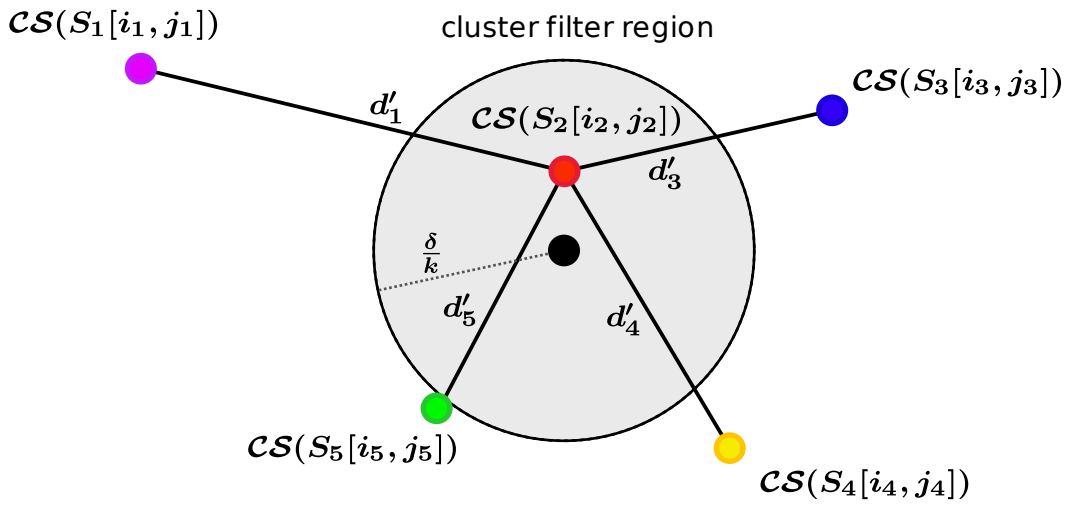
- Suchraum: $\mathcal{O}(n^{2k})$ Kombinationen von Teilworten von $S_1, ..., S_k$
- Cluster-Filter:



- Median-Distanzschwellwert: $\sum_{\ell=1}^{k} d_\ell \leq \delta$

Variante 2: Median-Gencluster — Reduktion des Suchraums

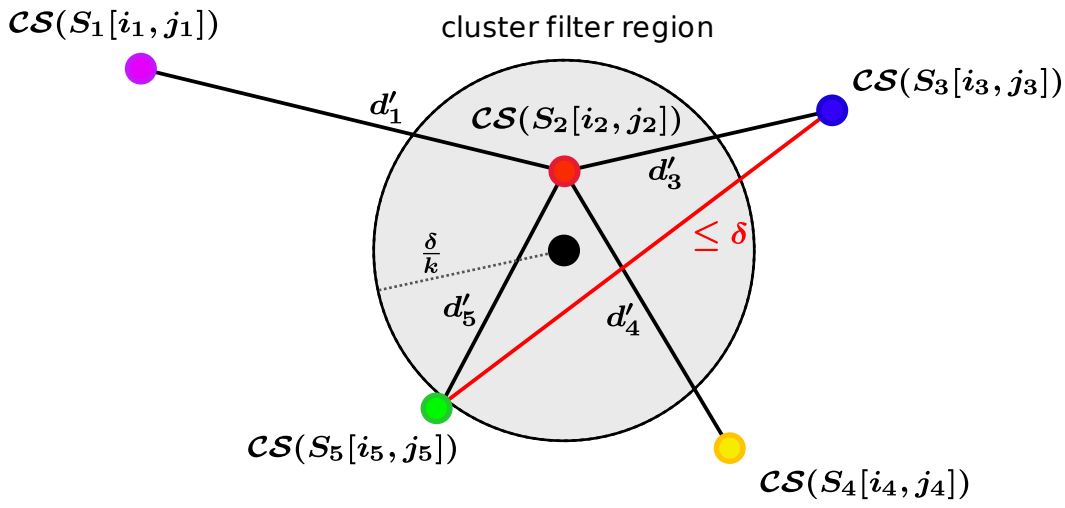
- Suchraum: $\mathcal{O}(n^{2k})$ Kombinationen von Teilworten von $S_1, ..., S_k$
- Cluster-Filter:



- Median-Distanzschwellwert: $\sum_{\ell=1}^{k} d_\ell \leq \delta$
- Cluster-Filter-Distanzschwellwert $\sum_{\ell=1}^{k} d'_\ell \leq 2\frac{k-1}{k}\delta$

Variante 2: Median-Gencluster — Reduktion des Suchraums

- Suchraum: $\mathcal{O}(n^{2k})$ Kombinationen von Teilworten von $S_1, ..., S_k$
- Cluster-Filter:



- Median-Distanzschwellwert: $\sum_{\ell=1}^{k} d_\ell \leq \delta$
- Cluster-Filter-Distanzschwellwert $\sum_{\ell=1}^{k} d'_\ell \leq 2\frac{k-1}{k}\delta$
- Paarweiser Distanz-Schwellwert: $d(S_\ell(i_\ell, j_\ell), S_m(i_m, j_m)) \leq \delta$

3-Schritt-Algorithmus zum Finden von Median-Genclustern

- Schritt 1: Berechne alle Cluster-Filter für $S_1, \ldots, S_k$
- Schritt 2: Berechne für jeden Cluster-Filter C alle Kombinationen mit Teilworten aus den anderen Sequenzen, für die:

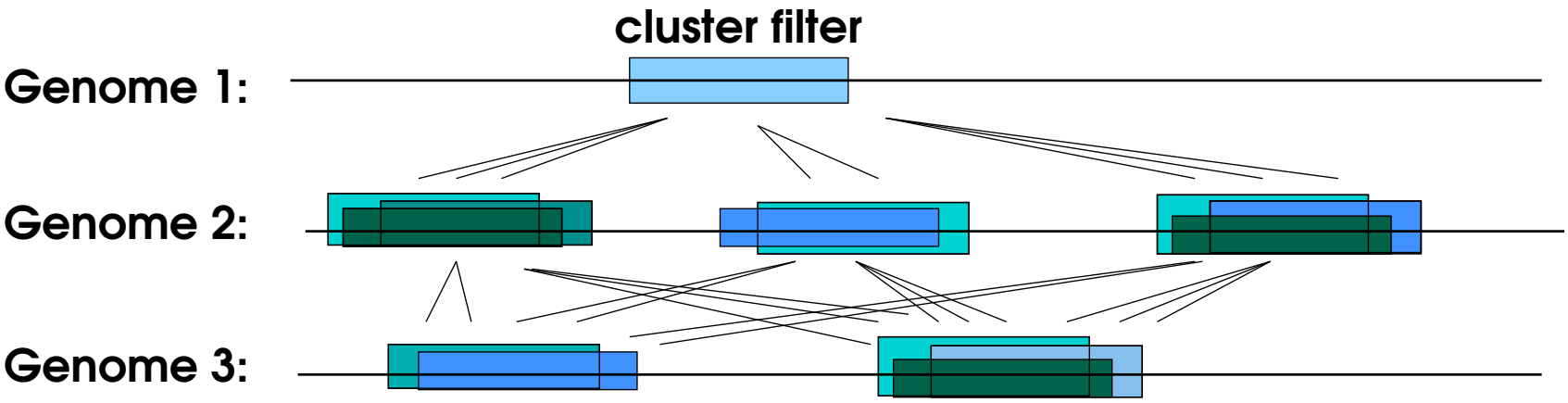
$$\sum_{\ell=1}^{k} D\big(C, \mathcal{CS}(S_\ell[i_\ell, j_\ell])\big) \;\leq\; 2\frac{k-1}{k}\delta$$

- Schritt 3:
  - Berechne den Median/die Mediane für jedes k-Tupel (aus Schritt 2)
  - Vergleiche die Median-Distanz mit dem gegebenen Distanz-Schwellwert δ

Schritt 1: Berechnung der Cluster-Filter für $S_1, \ldots, S_k$

- Jeder Cluster-Filter für einen gegebenen Distanz-Schwellwert δ ist ein Referenz-Gencluster für $2\frac{k-1}{k}\delta$
- Verwende den Algorithmus zur Berechnung der Referenz-Gencluster
- Zeitkomplexität: $\mathcal{O}(k^2 n^2 (1 + \delta^2))$

Schritt 2: Aufzählung aller k-Tupel



Idee: Ausgehend von jedem Cluster-Filter C, zähle alle Kombinationen mit Teilworten in den anderen Sequenzen auf, für die C ein Cluster-Filter ist

- mögliche kombinatorische Explosion: $\mathcal{O}(n^{2k})$
- bis zu $\mathcal{O}(\delta^{2k})$ Varianten eines k-Tupels

Schritt 3: Berechnung des Medians für jedes k-Tupel

- Mehrheitsentscheid: $\mathcal{O}(k|\Sigma|)$ Zeit und $\mathcal{O}(k|\Sigma|)$ Platz
- Vergleich der Gesamtdistanz mit dem Schwellwert δ

	1	2	3	4	5	6	7	8	9
$\mathcal{CS}(S_1[i_1, j_1])$	1	1	1	1	0	1	0	1	0
$\mathcal{CS}(S_2[i_2, j_2])$	1	1	0	1	0	1	1	1	1
$\mathcal{CS}(S_3[i_3, j_3])$	1	1	1	1	1	1	0	1	1
$\mathcal{CS}(S_4[i_4, j_4])$	1	1	1	1	1	1	1	1	0
$\mathcal{CS}(S_5[i_5, j_5])$	1	0	1	1	0	1	0	1	1
$\sum$	5	4	4	5	2	5	2	5	3

Schritt 3: Berechnung des Medians für jedes k-Tupel

- Mehrheitsentscheid: $\mathcal{O}(k|\Sigma|)$ Zeit und $\mathcal{O}(k|\Sigma|)$ Platz
- Vergleich der Gesamtdistanz mit dem Schwellwert δ

	1	2	3	4	5	6	7	8	9
$\mathcal{CS}(S_1[i_1, j_1])$	1	1	1	1	0	1	0	1	0
$\mathcal{CS}(S_2[i_2, j_2])$	1	1	0	1	0	1	1	1	1
$\mathcal{CS}(S_3[i_3, j_3])$	1	1	1	1	1	1	0	1	1
$\mathcal{CS}(S_4[i_4, j_4])$	1	1	1	1	1	1	1	1	0
$\mathcal{CS}(S_5[i_5, j_5])$	1	0	1	1	0	1	0	1	1
$\sum$	5	4	4	5	2	5	2	5	3

$$\text{median} = \{1, 2, 3, 4, 6, 8, 9\}$$

Variante 2: Median-Gencluster — Zusammenfassung

- Optimalitätskriterium (+)
- Filter-Ansatz und weitere Optimierungs-Schritte erlauben Reduktion des Suchraums (+)
- Suchraum wächst exponentiell mit der Anzahl Genome (−)
- In der Praxis: anwendbar auf mehrere Genome (+)
- δ muss mit der Anzahl untersuchter Genome wachsen
  $\longrightarrow$ Problem mit Vorkommen einzigartiger Gene (−)

Überblick

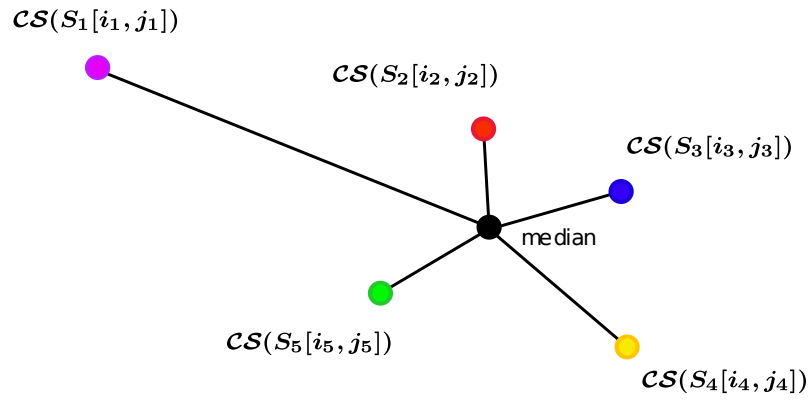
1. Komparative Genomanalyse
2. Approximative Gencluster
   - Referenz-Gencluster
   - Median-Gencluster
   - Center-Gencluster
3. Experimentelle Ergebnisse
4. Rekonstruktion phylogenetischer Bäume
5. Zusammenfassung und Ausblick

Variante 3: Center-Gencluster

- Distanzen zum „Konsensus-Gencluster“ sollten ähnlich sein

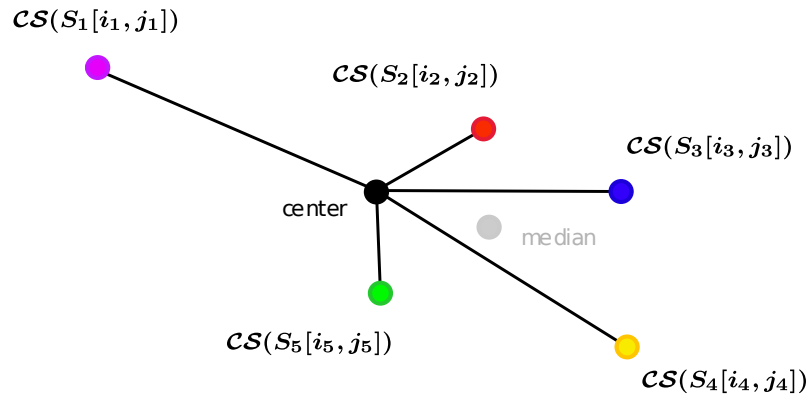
Variante 3: Center-Gencluster

- Distanzen zum „Konsensus-Gencluster“ sollten ähnlich sein
- Median: minimiert die Gesamt-Distanz zum Konsensus



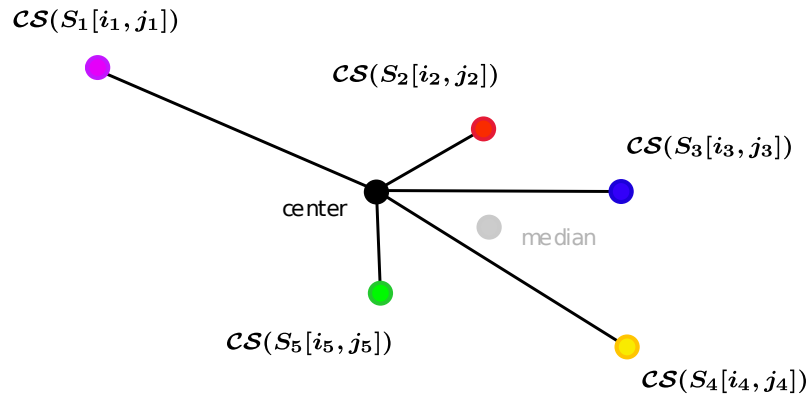
Variante 3: Center-Gencluster

- Distanzen zum „Konsensus-Gencluster“ sollten ähnlich sein
- Median: minimiert die Gesamt-Distanz zum Konsensus
- Center: minimiert die maximale paarweise Distanz



Variante 3: Center-Gencluster

- Distanzen zum „Konsensus-Gencluster“ sollten ähnlich sein
- Median: minimiert die Gesamt-Distanz zum Konsensus
- Center: minimiert die maximale paarweise Distanz



- Definition: $M \subseteq \Sigma$ ist ein Center von k Zeichenmengen $C_1, \ldots, C_k$ gdw. für jedes $M' \subseteq \Sigma$ gilt:

$$\max_{1 \leq \ell \leq k} (d(M, C_\ell)) \leq \max_{1 \leq \ell \leq k} (d(M', C_\ell))$$

Variante 3: Center-Gencluster — Problemdefinition

Gegeben:

- Sequenzen $S_1, \ldots, S_k$ über dem Alphabet der Gene Σ
- s (minimale Clustergröße)
- $\delta_{pw} \approx \frac{\delta}{k}$ (paarweiser Distanz-Schwellwert)

Gesucht: alle $M \subseteq \Sigma$ mit

- M ist Center für ein $S_1[i_1, j_1], ..., S_k[i_k, j_k]$
- $D(M, \mathcal{CS}(S[i_\ell, j_\ell])) \leq \delta_{pw}$ für alle $1 \leq \ell \leq k$
- $|M| \geq s$

Eine solche Menge M heißt Center-Gencluster von $S_1, \ldots, S_k$.

Modifizierter Algorithmus für die Center-Berechnung

- Schritt 1: Berechnung der Cluster-Filter
  - nur in einer Sequenz
  - stärkere Einschränkung durch die paarweise Distanz: $\delta \rightarrow 2\frac{\delta}{k}$

$$\mathcal{O}(k^2 n^2(1+\delta^2)) \rightarrow \mathcal{O}(kn^2(1+(\tfrac{\delta}{k})^2))$$

- Schritt 2: Kombination jedes Cluster-Filters C mit Teilworten aus den anderen Sequenzen, für die:

$$\sum_{\ell=1}^{k} D(C, \mathcal{CS}(S_\ell[i_\ell, j_\ell])) \leq 2\frac{k-1}{k}\delta$$

  - mögliche kombinatorische Explosion: $\mathcal{O}(n^k)$
  - realistischer: $\mathcal{O}(\delta^k) \rightarrow \mathcal{O}((\frac{\delta}{k})^k)$
- Schritt 3: Filter k-Tupel mit Median-Bedingung $\mathcal{O}(k|\Sigma|)$
- Schritt 4: Berechnung des Centers für die verbleibenden k-Tupel $\mathcal{O}(2^\sigma)$ (σ = lokale Alphabetgröße)

Variante 3: Center-Gencluster — Zusammenfassung

- Vernünftigeres Gencluster-Modell?
- Suchraum wächst exponentiell mit der Anzahl Genome (−)
- Filter-Ansatz ist möglich (+)
- Algorithmus skaliert besser bei großer Anzahl Genome (+)
- Gencluster mit einer „Median-artigen“ Struktur sind schwerer auffindbar (−)

Überblick

1. Komparative Genomanalyse
2. Approximative Gencluster
3. Experimentelle Ergebnisse
4. Rekonstruktion phylogenetischer Bäume
5. Zusammenfassung und Ausblick

Experimentelle Ergebnisse

- Suche nach Genclustern in einem typischen Datensatz
- fünf γ-Proteobakterien:

Spezies-Name	refSeq-ID	# Gene
Buchnera aphidicola APS	NC_002528	564
Escherichia coli K12	NC_000913	4183
Haemophilus influenzae Rd	NC_000907	1709
Pasteurella multocida Pm70	NC_002663	2015
Xylella fastidiosa 9a5c	NC_002488	2680

Experimentelle Ergebnisse — Median-Gencluster

	$\delta = 0$	$\delta = 1$	$\delta = 5$	$\delta = 8$	$\delta = 10$
$s = 4$					
Laufzeit	7s	7s	28s	1h 39m	-
# Gencluster-Klassen	6	7	36	43	-
$s = 5$					
Laufzeit	7s	7s	9s	1m 7s	35h 40m
# Gencluster-Klassen	5	5	13	25	26
$s = 6$					
Laufzeit	7s	7s	8s	13s	2h 14m
# Gencluster-Klassen	3	3	6	17	17

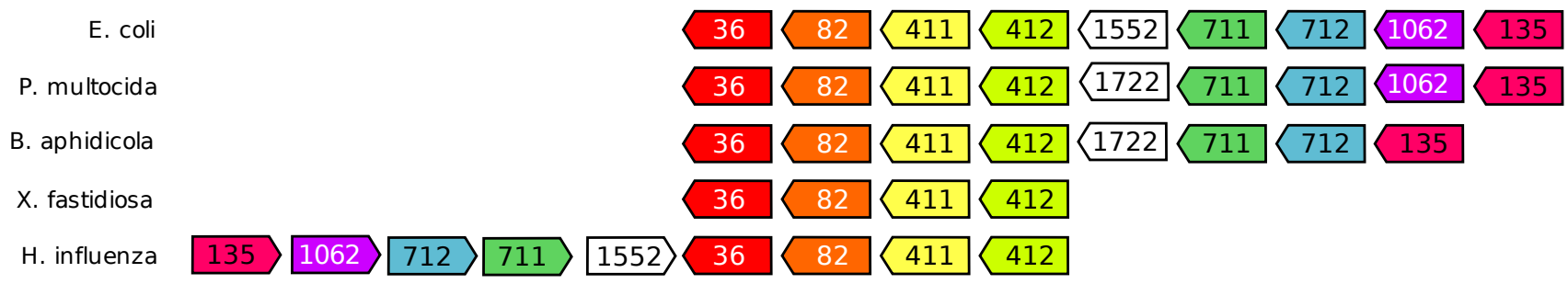
Experimentelle Ergebnisse — Center-Gencluster

	$\delta = 0$	$\delta = 1$	$\delta = 2$	$\delta = 3$	$\delta = 4$
$s = 4$					
Laufzeit	2s	4s	59m 16s	-	-
# Gencluster-Klassen	6	17	29	-	-
$s = 6$					
Laufzeit	2s	4s	32s	6h 0m	-
# Gencluster-Klassen	3	4	9	13	-
$s = 8$					
Laufzeit	2s	3s	6s	12m 51s	-
# Gencluster-Klassen	2	2	3	3	-
$s = 10$					
Laufzeit	2s	3s	5s	14s	3h 21m
# Gencluster-Klassen	1	1	2	2	3

Experimentelle Ergebnisse

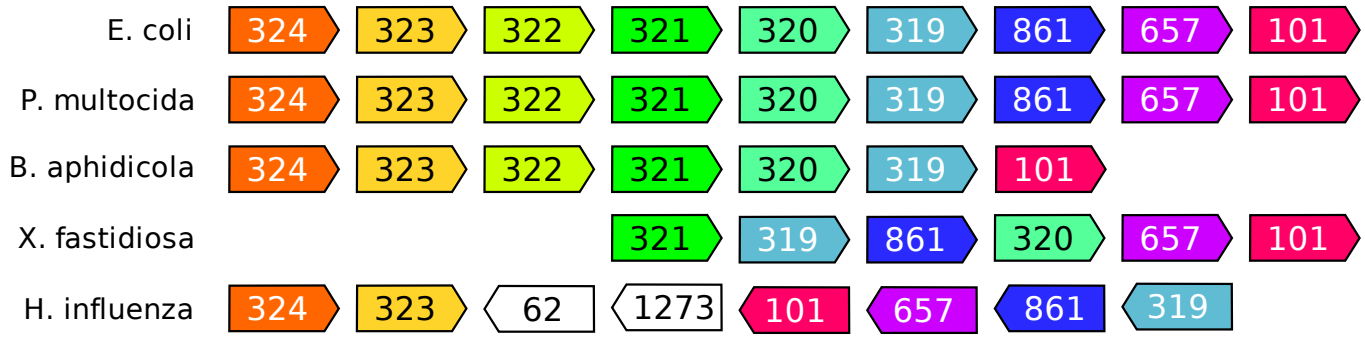
Zellteilung und Zellwand-Biosynthese:

Unklassifiziert:



Experimentelle Ergebnisse

Unklassifiziert:



Genom	Gen-ID	Annotation
E. coli	324	structural component; Ribosomal proteins - modification
	323	factor; Proteins - translation and modification (umk)
	322	enzyme; Central intermediary metabolism: interconversions (rrf)
	321	factor; Proteins - translation and modification (yaeM)
	320	1-deoxy-D-xylulose 5-phosphate reductoisomerase (yaeS)
	319	undecaprenyl pyrophosphate synthase
	861	enzyme; Fatty acid and phosphatidic acid biosynthesis (yaeL)
	657	zinc metallopeptidase (yzzY)
	101	conserved protein (ompH)

Überblick

1. Komparative Genomanalyse
2. Approximative Gencluster
3. Experimentelle Ergebnisse
4. Rekonstruktion phylogenetischer Bäume
5. Zusammenfassung und Ausblick

Phylogenetische Rekonstruktion aus kompletten Genomen

Idee: Paarweise Distanzen zwischen den Genomen, basierend auf

Grad der (noch) gemeinsam enthaltenen Gencluster

Phylogenetische Rekonstruktion aus kompletten Genomen

Idee: Paarweise Distanzen zwischen den Genomen, basierend auf

Grad der (noch) gemeinsam enthaltenen Gencluster

Klassische phylogenetische Analyse:

- Merkmalsbasiert: Vergleich heutiger Spezies/-eigenschaften

Phylogenetische Rekonstruktion aus kompletten Genomen

Idee: Paarweise Distanzen zwischen den Genomen, basierend auf

Grad der (noch) gemeinsam enthaltenen Gencluster

Klassische phylogenetische Analyse:

- Merkmalsbasiert: Vergleich heutiger Spezies/-eigenschaften
- Nicht: Rekonstruktion expliziter Rearrangement-Operationen (NP-schwer mit duplizierten Genen)

Phylogenetische Rekonstruktion aus kompletten Genomen

Idee: Paarweise Distanzen zwischen den Genomen, basierend auf

Grad der (noch) gemeinsam enthaltenen Gencluster

Klassische phylogenetische Analyse:

- Merkmalsbasiert: Vergleich heutiger Spezies/-eigenschaften
- Nicht: Rekonstruktion expliziter Rearrangement-Operationen (NP-schwer mit duplizierten Genen)
- Modellfrei: keine Annahmen über die zugrundeliegenden evolutionären Prozesse

Distanzmaße

- Zähle die Anzahl Intervalle in Genom S, deren Zeichenmenge ein approximatives Vorkommen in Genom T mit bis zu d Unterschieden hat:
- $CI(S, T, d) := |\{S[i, j] \text{ mit einem } d\text{-Ort in } T\}|$
- nicht symmetrisch: $CI(S, T, d) \neq CI(T, S, d)$
- Grundlegende Distanzformel:

  $$dist_1(S, T, d) = 1 - \frac{1}{2}\left(\frac{CI(S,T,d)}{CI(S,S,d)} + \frac{CI(T,S,d)}{CI(T,T,d)}\right)$$

- Alternative Distanzformel:

  $$dist_2(S, T, d) = 1 - \frac{1}{2}\left(\sqrt{\frac{CI(S,T,d)}{CI(S,S,d)}} + \sqrt{\frac{CI(T,S,d)}{CI(T,T,d)}}\right)$$

Methode

- Paarweiser Vergleich der Eingabe-Genome bezüglich der Anzahl approximativer Gencluster
- Berechnung einer Distanzmatrix
- Berechnung eines phylogenetischen Baums (Algorithmus von Fitch-Margoliash)
- Vergleich mit dem Referenz-Baum (Robinson-Foulds-Metrik)

Experimentelle Ergebnisse

- „Benchmark-Datensatz“ für phylogenetische Rekonstruktionen auf Basis ganzer Genome
- 12 γ-Proteobakterien

Abkürzung	Spezies-Name	# Gene
BAPHI	*Buchnera aphidicola APS*	564
ECOLI	*Escherichia coli K12*	4183
HAEIN	*Haemophilus influenzae Rd*	1709
PAERU	*Pseudomonas aeruginosa PA01*	5540
PMULT	*Pasteurella multocida Pm70*	2015
SALTY	*Salmonella typhimurium LT2*	4203
WGLOS	*Wigglesworthia glossinidia brevipalpis*	653
XAXON	*Xanthomonas axonopodis pv. citri 306*	4192
XCAMP	*Xanthomonas campestris*	4029
XFAST	*Xylella fastidiosa 9a5c*	2680
YPEST-CO92	*Yersinia pestis CO_92*	3599
YPEST-KIM	*Yersinia pestis KIM5 P12*	3879

Experimentelle Ergebnisse

- Phylogenetische Bäume, berechnet für alle Kombinationen von:
  - zwei Distanzformeln
  - vier Maße der Konserviertheit
  - fünf Distanz-Schwellwerte d
- Robinson-Foulds-Distanzen zum Referenz-Baum:

	$dist_1$					$dist_2$				
$d =$	0	1	5	10	20	0	1	5	10	20
CI	4	4	2	0	0	2	2	2	2	0
CI_{size}	12	8	6	6	6	8	6	0	0	0
CI_{deg}	4	4	4	2	0	2	2	2	2	2
$CI_{size,deg}$	12	8	6	6	6	8	6	0	0	0

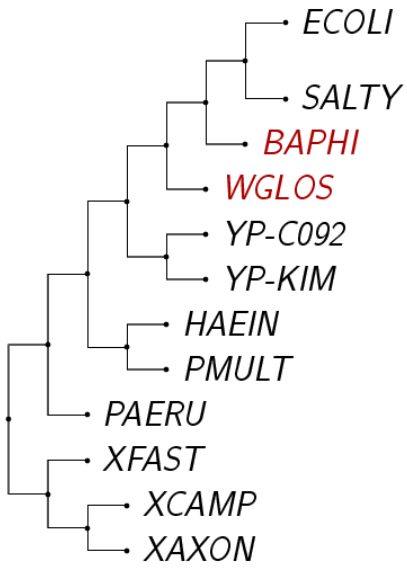
Experimentelle Ergebnisse

- Phylogenetische Bäume, berechnet für alle Kombinationen von:
  - zwei Distanzformeln
  - vier Maße der Konserviertheit
  - fünf Distanz-Schwellwerte d
- Robinson-Foulds-Distanzen zum Referenz-Baum:

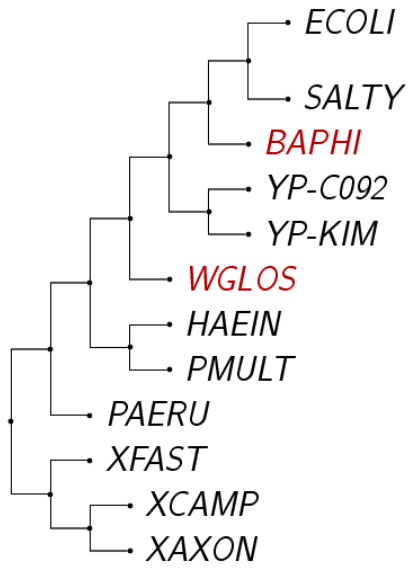
	$dist_1$					$dist_2$				
$d =$	0	1	5	10	20	0	1	5	10	20
CI	4	4	2	0	0	2	2	2	2	0
CI_{size}	12	8	6	6	6	8	6	0	0	0
CI_{deg}	4	4	4	2	0	2	2	2	2	2
$CI_{size,deg}$	12	8	6	6	6	8	6	0	0	0

Vorhergesagte Bäume vs. Referenz-Baum

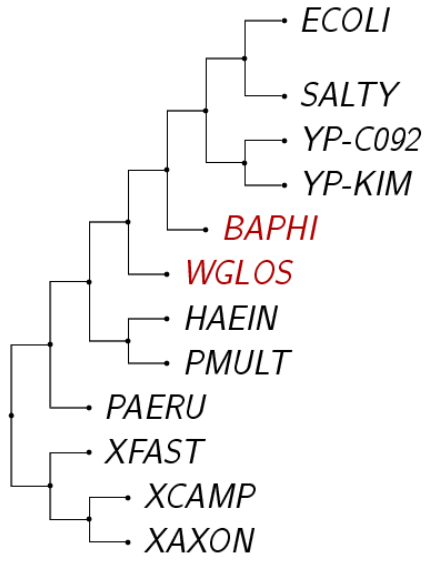
$T_{d=0}$



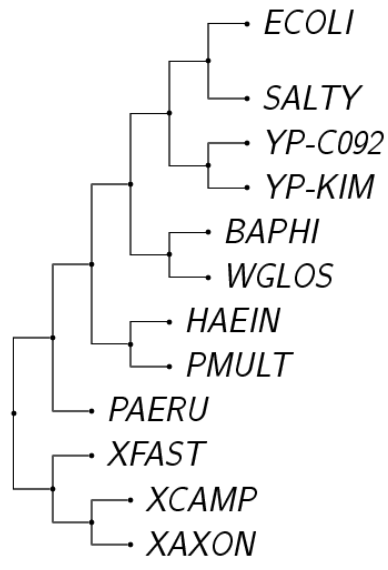
$T_{d=1}$



$T_{d=5}$



$T_{d=10} = T_{ref}$



Zusammenfassung - Phylogenetische Rekonstruktion

- Anwendung von approximativen Genclustern
- Schnelle Berechnung der Distanzen (+)
- Gute Qualität der vorhergesagten Baum-Topologien (+)
- Besser als andere merkmalsbasierte Methoden (Breakpoint-Distanz, gemeinsame Intervalle) (+)
- Unklare Interpretation der ermittelten Astlängen (−)

Überblick

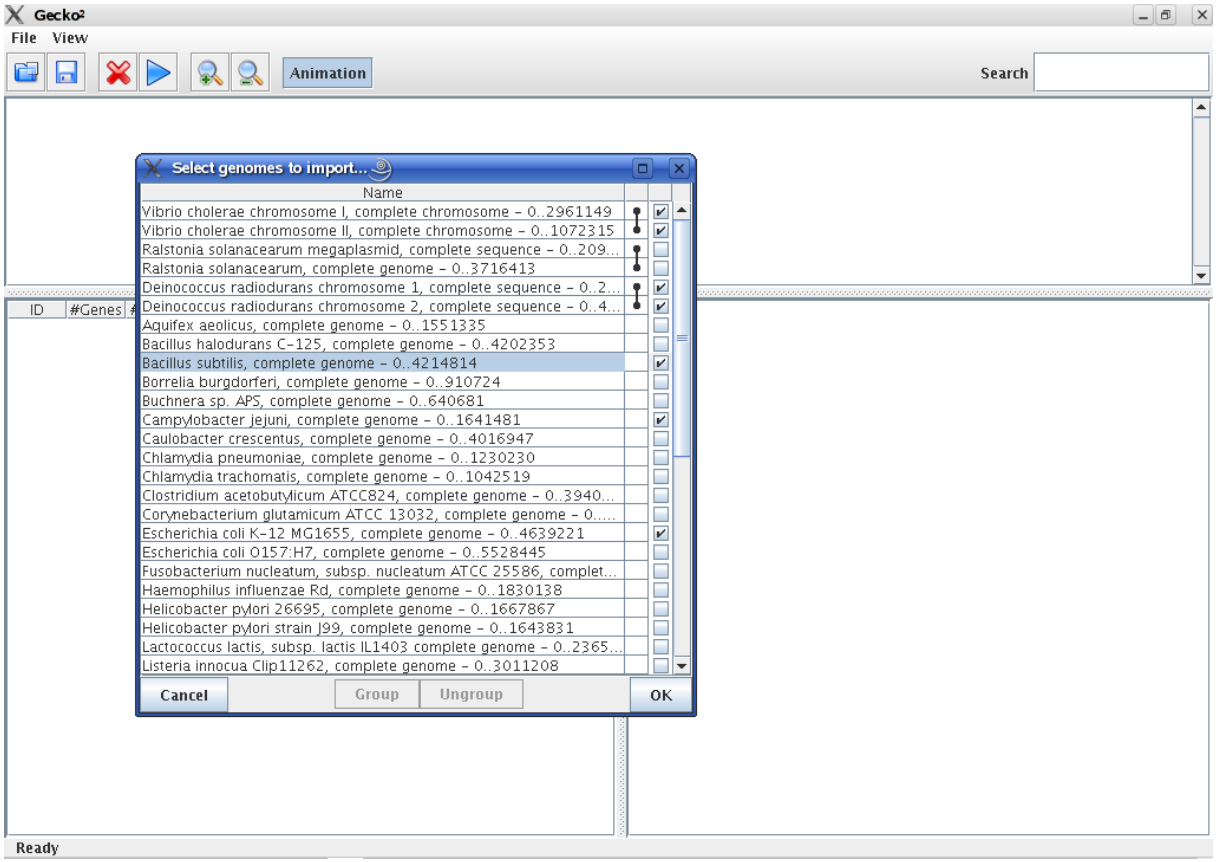
1. Komparative Genomanalyse
2. Approximative Gencluster
3. Experimentelle Ergebnisse
4. Rekonstruktion phylogenetischer Bäume
5. Zusammenfassung und Ausblick

Zusammenfassung

- Approximative Gencluster und ihre Anwendungen
- Software: Gecko 2

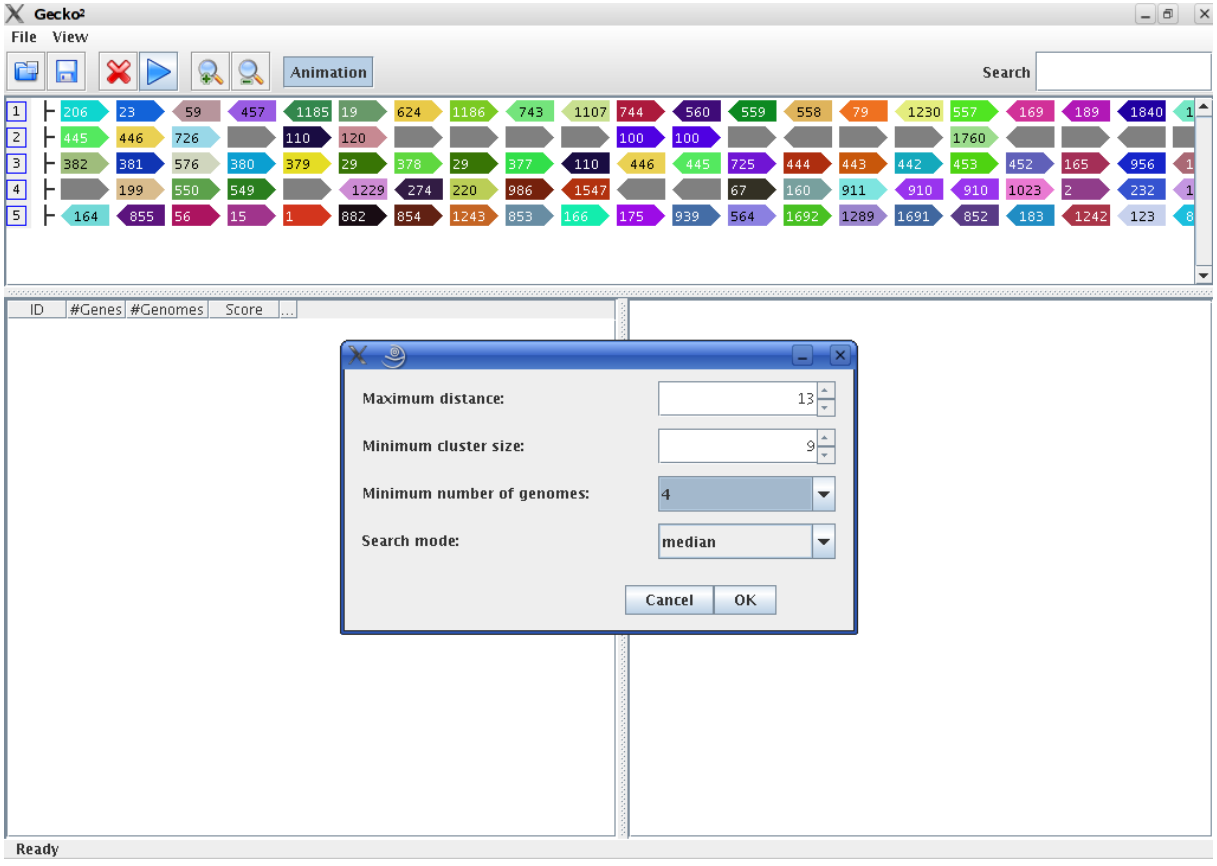
Zusammenfassung

- Approximative Gencluster und ihre Anwendungen
- Software: Gecko 2



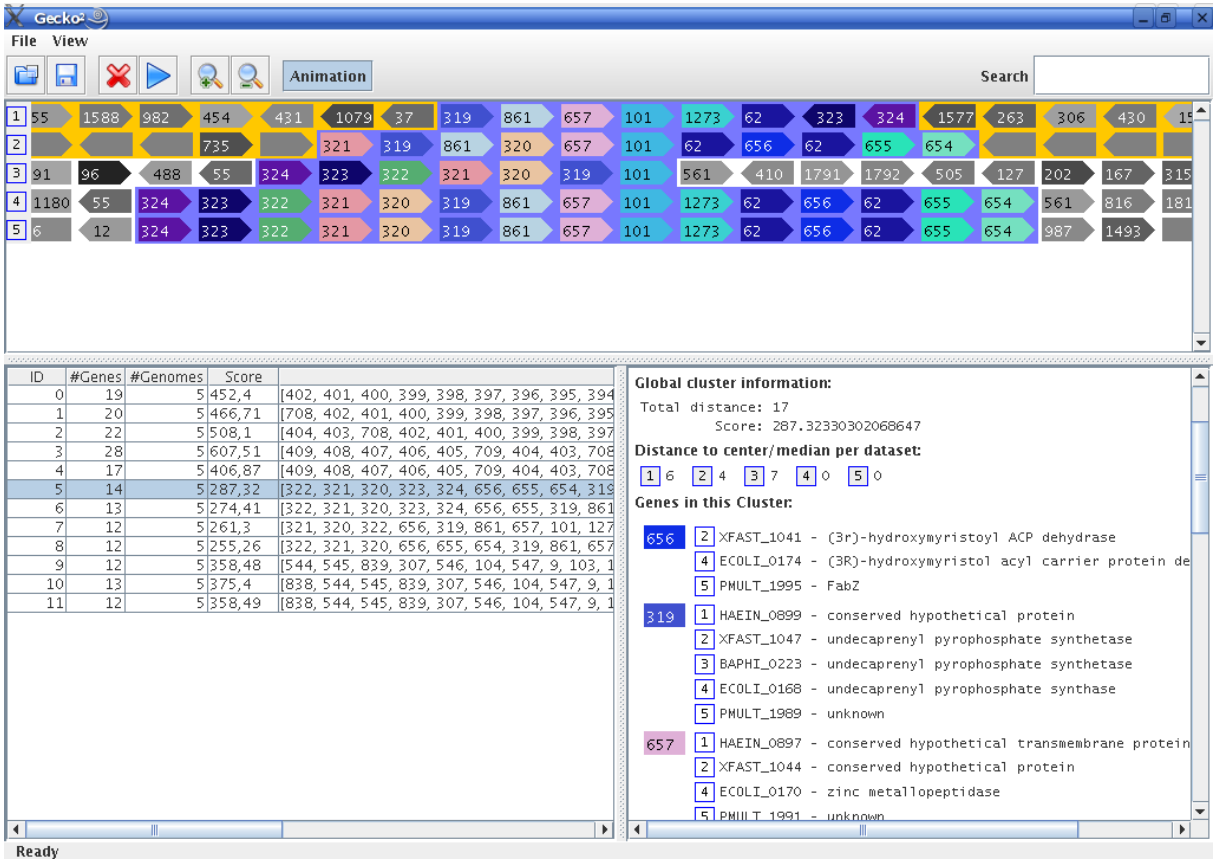
Zusammenfassung

- Approximative Gencluster und ihre Anwendungen
- Software: Gecko 2



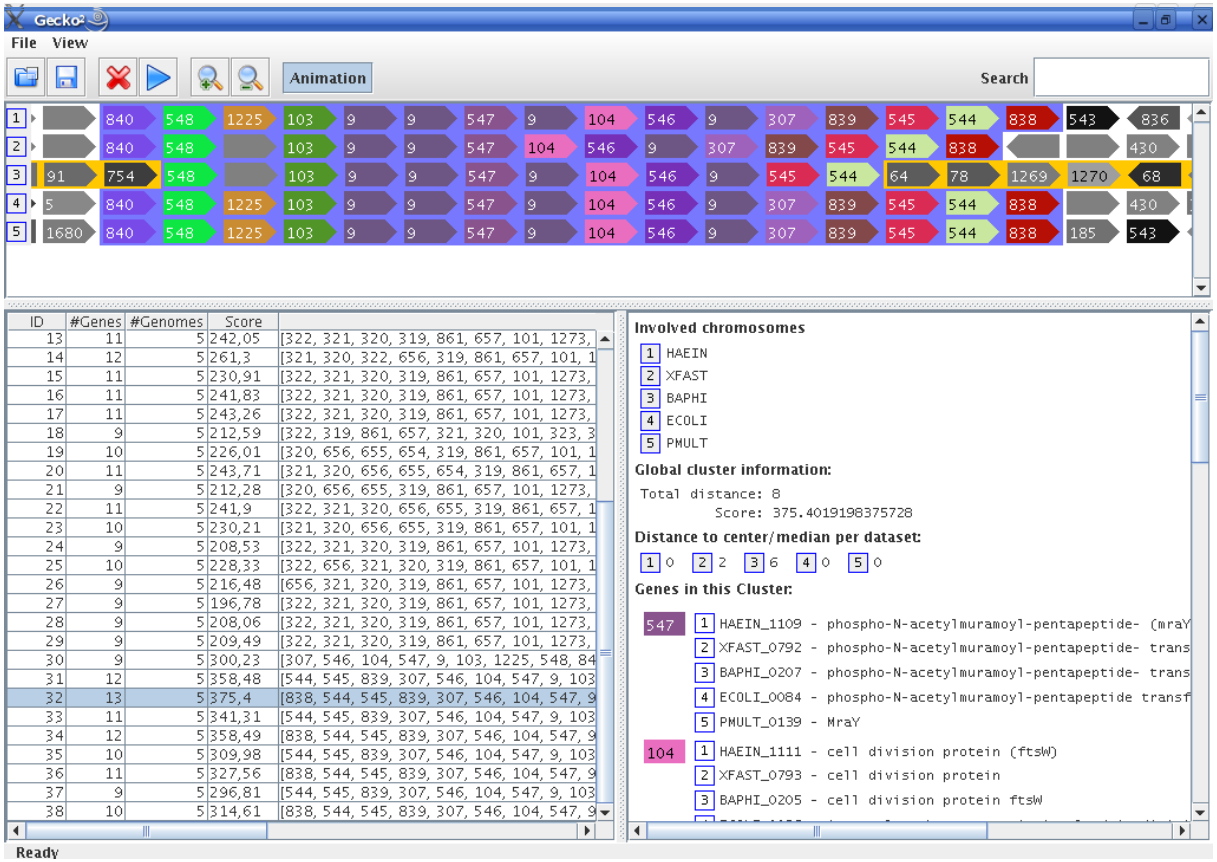
Zusammenfassung

- Approximative Gencluster und ihre Anwendungen
- Software: Gecko 2



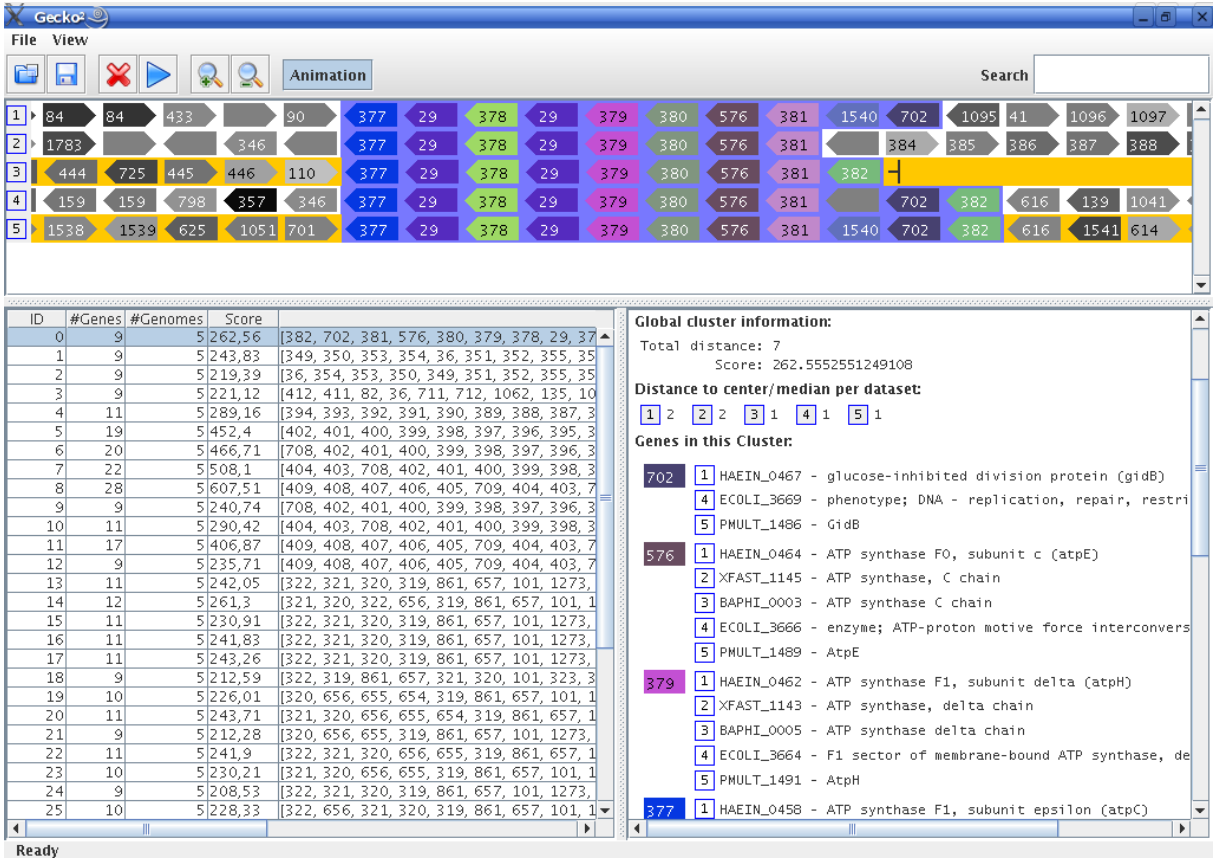
Zusammenfassung

- Approximative Gencluster und ihre Anwendungen
- Software: Gecko 2



Zusammenfassung

- Approximative Gencluster und ihre Anwendungen
- Software: Gecko 2



Ausblick

- Verbesserung der praktischen Laufzeiten
- Hinzufügen der evolutionären Unterschiede innerhalb der Gencluster, um Consensus-Mengen zu verbessern
- Unabhängigkeit von Homologie-Zuweisungen
  $\longrightarrow$ Sequenzähnlichkeiten der Gene statt feste Gruppierung
- Anwendung der Algorithmen auf andere Sequenztypen:
  $\longrightarrow$ Sequenzen von Transkriptionsfaktor-Bindungsstellen
  $\longrightarrow$ Sequenzen von Worten (Texte)

Dank an:

- **Katharina Jahn**
- Léon Kuchenbecker
- Sebastian Böcker (FSU Jena)
- Julia Mixtacki

Dank an:

- **Katharina Jahn**
- Léon Kuchenbecker
- Sebastian Böcker (FSU Jena)
- Julia Mixtacki

... und für Ihre Aufmerksamkeit!