

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

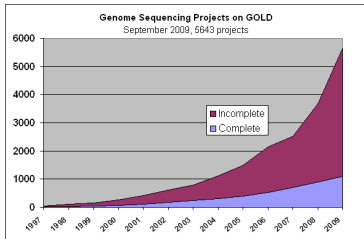
- 1 Komparative Genomanalyse
- 2 Approximative Gencluster
- 3 Experimentelle Ergebnisse
- 4 Rekonstruktion phylogenetischer Bäume
- 5 Zusammenfassung und Ausblick

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: ...cgtaggctacgccctaggcttcagtcgtattgatactgtagttgcttacgtagcatgatcagctgctgagtcgtacg...

genome 2: ...cgtacagctacgtcaacggttcacgtattgatgccctcgtagtcacgctacgtacgtaatgctgagactcatcgtacg...

genome 3: ...cgtagatgagctaaagtcgtattgatactcggttgagtacgtaggtacatgatgtgctaagagactgtcgtcgtacg...

Genom-Modell

genome 1: ...cgtaggctacgcc taggctcagtcgtattgatactgtagttgcttacgtagcatgatcagctgctgagtcgtac...

genome 2: ...cgtacagctacgtcaacggttcacgtattgatgccctcgtagtcacgctacgtacgtaatgctgagactcatcgtacg...

genome 3: ...cgtagatgagctaaagtcgtattgatactcggttgagtacgtaggatcatgatgtgctaagagactgtcgtcgtacg...



genome 1:



genome 2:



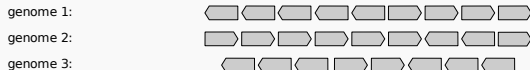
genome 3:



sequence of genes

Genom-Modell

genome 1: ...cgtaggctacgcc taggcttcagtcgattgatactgtagttgcttacgtagcatgatcagctctgagtcgtacg...
genome 2: ...cgtacagctacgtcaacggttcacgtattgatgccctcgtagtcacgctacgtacgtaatgctgagactcatcgtacg...
genome 3: ...cgtagatgagctaaagtcgattgatactcggttgagtacgtaggtacatgatgtgctaaagagactgtcgtcgtacg...



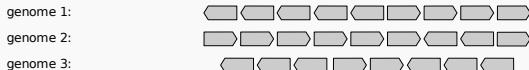
sequence of genes



homology based
labeling of genes

Genom-Modell

genome 1: ...cgtaggctacgcc taggcttcagtcgattgatactgtagttgcttacgtagcatgatcagctgctgagtcgtacg...
genome 2: ...cgtacagctacgtcaacggttcacgattgatgcccttcgtagtcacgctacgtacgtaatgctgagactcatcgtacg...
genome 3: ...cgtagatgagctaaagtcgattgatactcggttgagtacgtaggtacatgatgtgctaagagactgtcgtcgtacg...



sequence of genes



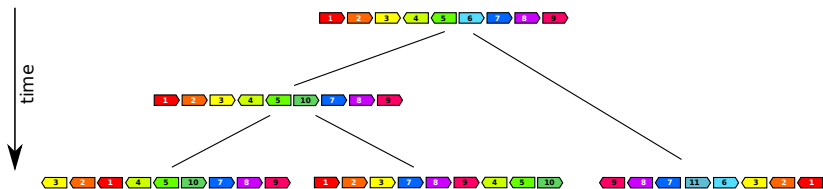
homology based
labeling of genes



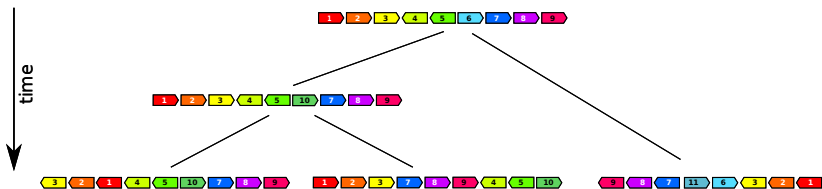
genome 1: 3 2 1 4 5 10 7 8 9
genome 2: 1 2 3 7 8 9 4 5 10
genome 3: 9 8 7 11 6 3 2 1

sequence of unsigned
integers

Evolution kompletter Genome



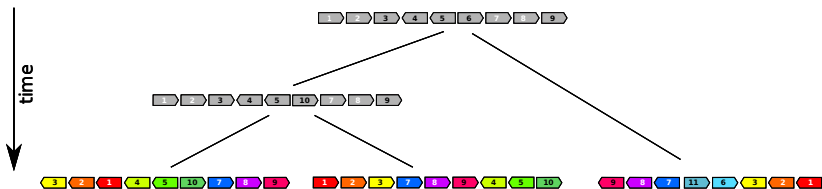
Evolution kompletter Genome



Zwei komplementäre Kräfte beeinflussen die Genreihenfolge:

- zufällige Umordnungen → phylogenetische Beziehungen
- Selektionsdruck → funktionelle Beziehungen

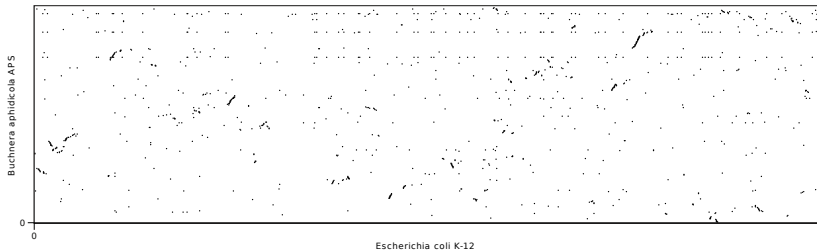
Evolution kompletter Genome



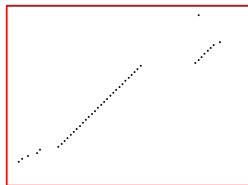
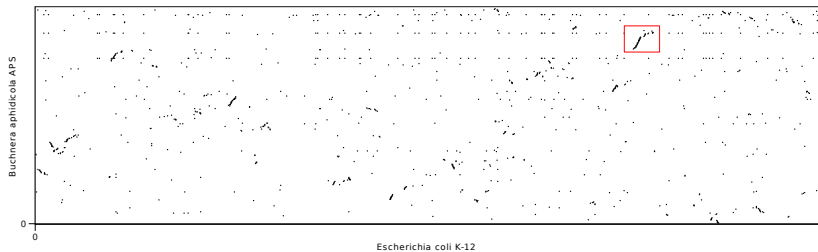
Zwei komplementäre Kräfte beeinflussen die Genreihenfolge:

- zufällige Umordnungen → phylogenetische Beziehungen
- Selektionsdruck → funktionelle Beziehungen

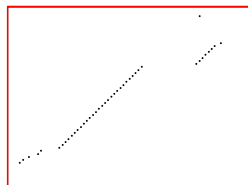
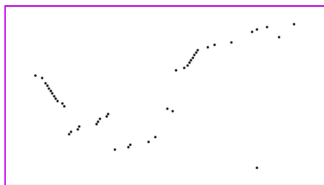
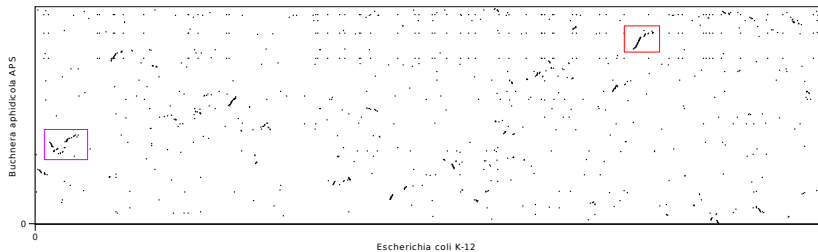
Dot-Plot der Genreihenfolge



Dot-Plot der Genreihenfolge



Dot-Plot der Genreihenfolge



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

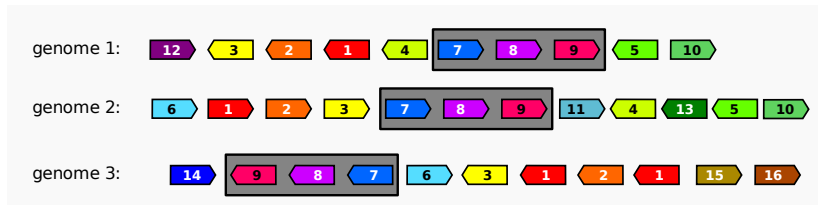
Gencluster

genome 1: 

genome 2: 

genome 3: 

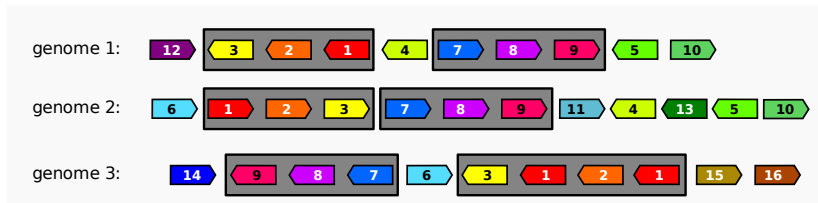
Gencluster



- **Kollineare Gencluster:**

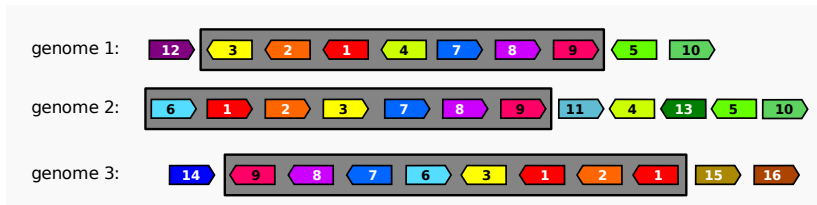
- Menge der Gene liegen *en bloc* in mehreren Genomen.
- Gen-Reihenfolge ist komplett erhalten (abgesehen von Inversionen).

Gencluster



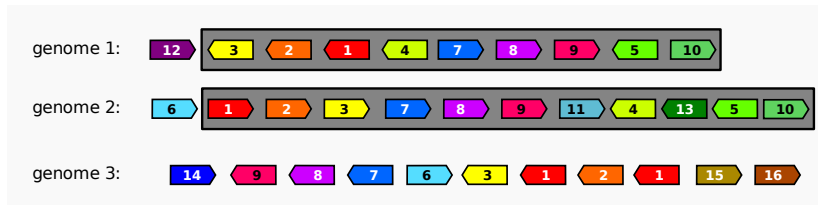
- **Kollineare Gencluster:**
 - Menge der Gene liegen *en bloc* in mehreren Genomen.
 - Gen-Reihenfolge ist komplett erhalten (abgesehen von Inversionen).
- **Perfekte Gencluster (*common intervals*):**
 - Genreihenfolge innerhalb des Block wird vernachlässigt.
 - Multiple Vorkommen eines Gens im Cluster 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, \dots, S_k über Alphabet Σ
- $n =$ maximale Länge der S_1, \dots, 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:**

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

S_1 : 0 | 2 5 13 7 1 3 5 4 6 9 8 7 3 10 4 11 12 5 | 0

S_2 : 0 4 6 8 9 3 14 1 9 6 8 4 1 10 2 15 | 0

S_3 : 0 | 7 17 10 1 16 18 9 8 6 4 6 19 11 5 20 3 | 0

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:** $([i_1, j_1], \dots, [i_k, j_k])$
mit einer Menge $C \subseteq \Sigma$ für einen **Schwellwert** δ :

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

S_1 : 0 | 2 5 13 7 1 3 5 4 6 9 8 7 3 10 4 11 12 5 | 0

S_2 : 0 | 4 6 8 9 3 14 1 9 6 8 4 1 10 2 15 | 0

S_3 : 0 | 7 17 10 1 16 18 9 8 6 4 6 19 11 5 20 3 | 0

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

Eine solche Menge M heißt **Referenz-Gencluster** von S_1, \dots, S_k .

Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
→ 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

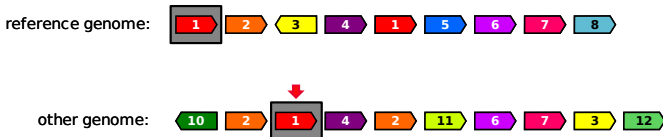
- Perfektes Vorkommen in mindestens einem Genom
→ **Startpunkt der Berechnung**
- Naiver Ansatz: berechne für alle Teilworte die paarweisen Distanzen zu allen anderen Teilworten $\mathcal{O}(k^2 n^4)$
- Unser Ansatz: Erweiterung des Algorithmus *Connecting Intervals* (Schmidt & JS, 2004)
- Beispiel: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

reference genome: 

other genome: 

Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
→ 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: $s = 6$, $\delta = 2$



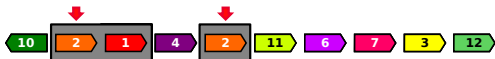
Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
→ **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: $s = 6, \delta = 2$

reference genome:



other genome:



Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
→ **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: $s = 6, \delta = 2$

reference genome:



other genome:



Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
→ 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: $s = 6$, $\delta = 2$

reference genome:



other genome:



Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
→ 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: $s = 6$, $\delta = 2$

reference genome:



other genome:



Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
→ **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: $s = 6$, $\delta = 2$

reference genome:



other genome:



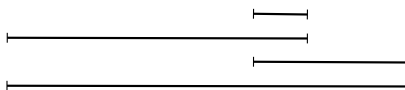
Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
→ **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: $s = 6$, $\delta = 2$

reference genome:



other genome:



superintervals with up to 2 inserted genes

0 inserted / 5 missing

1 inserted / 2 missing

1 inserted / 4 missing

2 inserted / 1 missing

Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
→ **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: $s = 6$, $\delta = 2$

reference genome:



other genome:



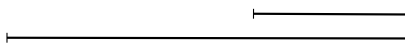
Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
→ 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: $s = 6$, $\delta = 2$

reference genome:



other genome:



superintervals with up to 2 inserted genes

0 inserted / 4 missing

1 inserted / 1 missing

Variante 1: Referenz-Gencluster

- Perfektes Vorkommen in mindestens einem Genom
→ **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: $s = 6$, $\delta = 2$

reference genome:



other genome:



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

- 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?
→ 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
→ Finde eine Konsensus-Menge, die die Abstände zu allen Vorkommen minimiert.
- Kein perfektes Vorkommen der Konsensus-Menge!
- Exponentieller Suchraum: $O(n^{2k})$ Teilwort-Kombinationen, $O(2^{|\Sigma|})$ mögliche Konsensus-Mengen

Variante 2: Median-Gencluster — Problemdefinition

Gegeben:

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

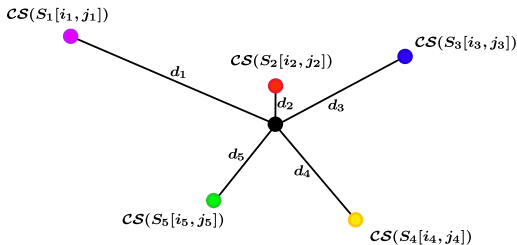
Gesucht: alle $M \subseteq \Sigma$ mit

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

Eine solche Menge M heißt **Median-Gencluster** von S_1, \dots, S_k .

Variante 2: Median-Gencluster — Reduktion des Suchraums

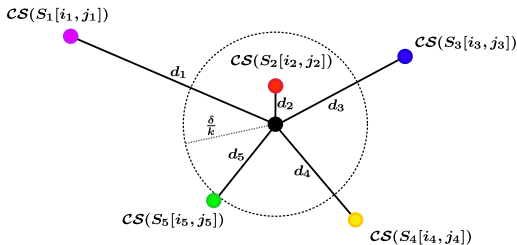
- Suchraum: $O(n^{2k})$ Kombinationen von Teilworten von S_1, \dots, S_k
- Cluster-Filter:



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

Variante 2: Median-Gencluster — Reduktion des Suchraums

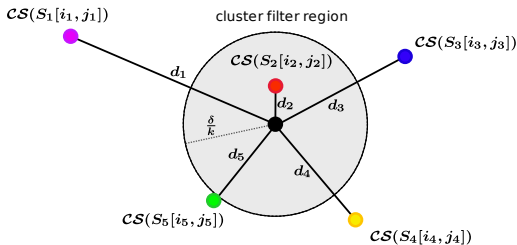
- Suchraum: $O(n^{2k})$ Kombinationen von Teilworten von S_1, \dots, S_k
- Cluster-Filter:



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

Variante 2: Median-Gencluster — Reduktion des Suchraums

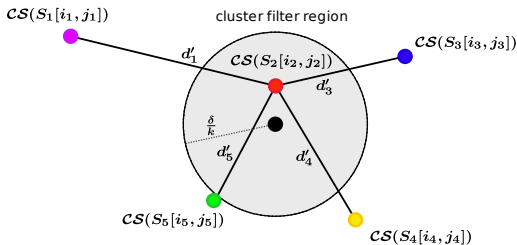
- Suchraum: $O(n^{2k})$ Kombinationen von Teilworten von S_1, \dots, S_k
- Cluster-Filter:



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

Variante 2: Median-Gencluster — Reduktion des Suchraums

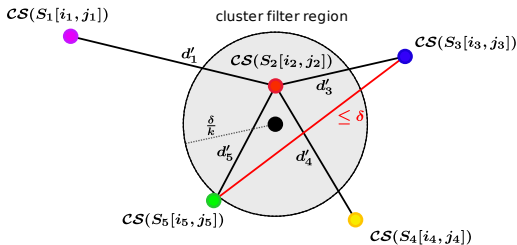
- Suchraum: $O(n^{2k})$ Kombinationen von Teilworten von S_1, \dots, S_k
- Cluster-Filter:



- Median-Distanzschwellewert: $\sum_{l=1}^k d_l \leq \delta$
- Cluster-Filter-Distanzschwellewert $\sum_{l=1}^k d'_l \leq 2 \frac{k-1}{k} \delta$

Variante 2: Median-Gencluster — Reduktion des Suchraums

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



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

3-Schritt-Algorithmus zum Finden von Median-Genclustern

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

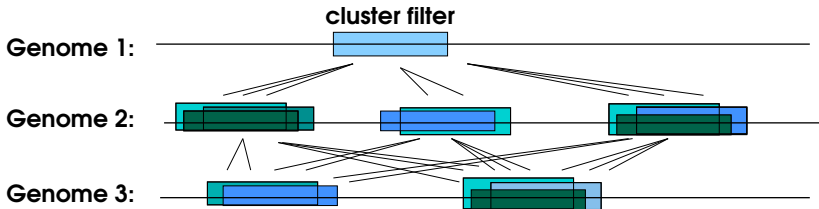
$$\sum_{\ell=1}^k D(C, \mathcal{CS}(S_\ell[i_\ell, j_\ell])) \leq 2 \frac{k-1}{k} \delta$$

- **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, \dots, 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: $O(n^{2k})$
- bis zu $O(\delta^{2k})$ 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
Σ	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
Σ	5	4	4	5	2	5	2	5	3

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
→ Problem mit Vorkommen einzigartiger Gene (-)

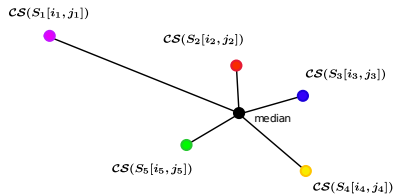
- 1 Komparative Genomanalyse
- 2 **Approximative Gencluster**
 - Referenz-Gencluster
 - Median-Gencluster
 - **Center-Gencluster**
- 3 Experimentelle Ergebnisse
- 4 Rekonstruktion phylogenetischer Bäume
- 5 Zusammenfassung und Ausblick

Variante 3: Center-Gencluster

- Distanzen zum „Konsensus-Gencluster“ sollten ähnlich sein

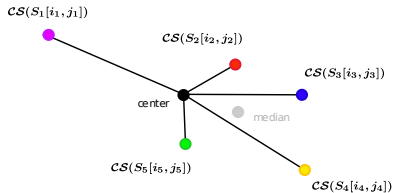
Variante 3: Center-Gencluster

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



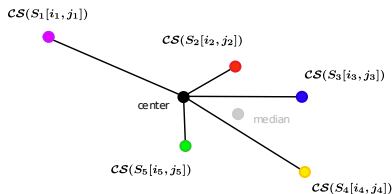
Variante 3: Center-Gencluster

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

Eine solche Menge M heißt **Center-Gencluster** von S_1, \dots, S_k .

Modifizierter Algorithmus für die Center-Berechnung

- **Schritt 1:** Berechnung der Cluster-Filter

- nur in einer Sequenz

- stärkere Einschränkung durch die paarweise Distanz: $\delta \rightarrow 2\frac{\delta}{k}$

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

- **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 $\mathcal{O}(2^\sigma)$ ($\sigma =$ 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
<i>Buchnera aphidicola</i> APS	NC_002528	564
<i>Escherichia coli</i> K12	NC_000913	4183
<i>Haemophilus influenzae</i> Rd	NC_000907	1709
<i>Pasteurella multocida</i> Pm70	NC_002663	2015
<i>Xylella fastidiosa</i> 9a5c	NC_002488	2680

Experimentelle Ergebnisse — Median-Gencluster

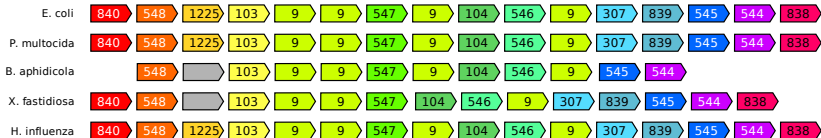
	$\delta = 0$	$\delta = 1$	$\delta = 5$	$\delta = 8$	$\delta = 10$
<hr/>					
$s = 4$					
Laufzeit	7s	7s	28s	1h 39m	-
# Gencluster-Klassen	6	7	36	43	-
<hr/>					
$s = 5$					
Laufzeit	7s	7s	9s	1m 7s	35h 40m
# Gencluster-Klassen	5	5	13	25	26
<hr/>					
$s = 6$					
Laufzeit	7s	7s	8s	13s	2h 14m
# Gencluster-Klassen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

Experimentelle Ergebnisse — Center-Gencluster

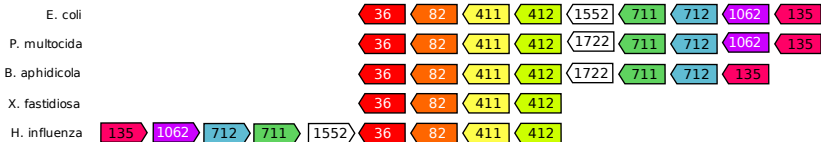
	$\delta = 0$	$\delta = 1$	$\delta = 2$	$\delta = 3$	$\delta = 4$
<hr/>					
$s = 4$					
Laufzeit	2s	4s	59m 16s	-	-
# Gencluster-Klassen	6	17	29	-	-
<hr/>					
$s = 6$					
Laufzeit	2s	4s	32s	6h 0m	-
# Gencluster-Klassen	3	4	9	13	-
<hr/>					
$s = 8$					
Laufzeit	2s	3s	6s	12m 51s	-
# Gencluster-Klassen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

Experimentelle Ergebnisse

Zellteilung und Zellwand-Biosynthese:

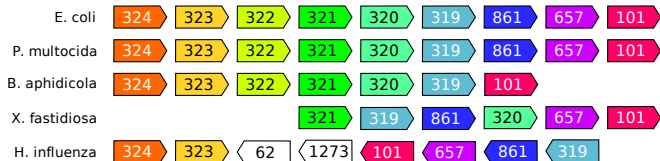


Unklassifiziert:



Experimentelle Ergebnisse

Unklassifiziert:



Genom	Gen-ID	Annotation
<i>E. coli</i>	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	<i>Buchnera aphidicola</i> APS	564
ECOLI	<i>Escherichia coli</i> K12	4183
HAEIN	<i>Haemophilus influenzae</i> Rd	1709
PAERU	<i>Pseudomonas aeruginosa</i> PA01	5540
PMULT	<i>Pasteurella multocida</i> Pm70	2015
SALTY	<i>Salmonella typhimurium</i> LT2	4203
WGLOS	<i>Wigglesworthia glossinidia</i> <i>brevipalpis</i>	653
XAXON	<i>Xanthomonas axonopodis</i> pv. <i>citri</i> 306	4192
XCAMP	<i>Xanthomonas campestris</i>	4029
XFAST	<i>Xylella fastidiosa</i> 9a5c	2680
YPEST-CO92	<i>Yersinia pestis</i> CO_92	3599
YPEST-KIM	<i>Yersinia pestis</i> 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

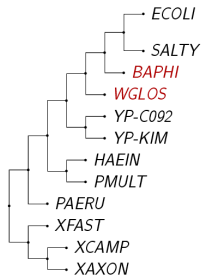
Experimentelle Ergebnisse

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

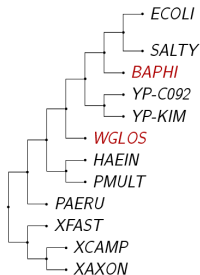
	$dist_1$					$dist_2$				
$d =$	0	1	5	10	20	0	1	5	10	20
CI	4	4	2	0	0	2	2	2	2	0
CI_{size}	12	8	6	6	6	8	6	0	0	0
CI_{deg}	4	4	4	2	0	2	2	2	2	2
$CI_{size,deg}$	12	8	6	6	6	8	6	0	0	0

Vorhergesagte Bäume vs. Referenz-Baum

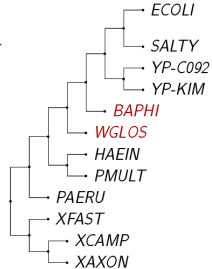
$T_{d=0}$



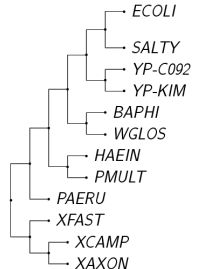
$T_{d=1}$



$T_{d=5}$



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



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

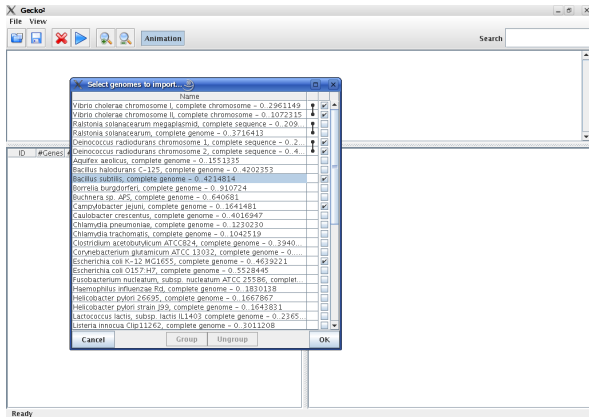
- 1 Komparative Genomanalyse
- 2 Approximative Gencluster
- 3 Experimentelle Ergebnisse
- 4 Rekonstruktion phylogenetischer Bäume
- 5 Zusammenfassung und Ausblick

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

The screenshot displays the Gecko 2 software interface. The main window shows a visualization of gene clusters across five genomes (labeled 1 to 5). Each genome is represented by a horizontal bar with colored segments representing genes. The colors correspond to different clusters. A search bar is visible in the top right corner.

A configuration dialog box is open in the foreground, allowing users to adjust the following parameters:

- Maximum distance: 13
- Minimum cluster size: 9
- Minimum number of genomes: 4
- Search mode: median

Buttons for "Cancel" and "OK" are located at the bottom of the dialog box. The status bar at the bottom left of the main window indicates "Ready".

Zusammenfassung

- Approximative Gencluster und ihre Anwendungen
- Software: Gecko 2

The screenshot displays the Gecko 2 software interface. The main window shows a genomic map with a grid of colored cells representing gene clusters. The cells are color-coded by cluster: 1 (yellow), 2 (orange), 3 (green), 4 (blue), 5 (purple), 6 (red), and 11 (grey). The grid contains numerical values representing gene counts for each cluster across different genomic regions.

ID	#Genes	#Genomes	Score
0	19	5	452,4
1	20	5	466,71
2	22	5	508,1
3	28	5	607,51
4	17	5	406,87
5	14	5	287,32
6	13	5	274,41
7	12	5	261,3
8	12	5	255,26
9	12	5	358,48
10	13	5	375,4
11	12	5	358,49

Global cluster information:
Total distance: 17
Score: 287,3233030208647

Distance to center/median per dataset:
1 | 6 | 2 | 4 | 3 | 7 | 4 | 0 | 5 | 0

Genes in this Cluster:

- 656 [2] XFAST_1041 - (3R)-hydroxytryptoyl ACP dehydrase
- [4] ECOLI_0174 - (3R)-hydroxytryptoyl acyl carrier protein de
- [5] PMULT_1995 - FabZ
- 319 [1] HAEIN_0899 - conserved hypothetical protein
- [2] XFAST_1047 - undecaprenyl pyrophosphate synthetase
- [3] BAPHI_0223 - undecaprenyl pyrophosphate synthetase
- [4] ECOLI_0168 - undecaprenyl pyrophosphate synthase
- [5] PMULT_1989 - unknown
- 657 [1] HAEIN_0897 - conserved hypothetical transmembrane protein
- [2] XFAST_1044 - conserved hypothetical protein
- [4] ECOLI_0170 - zinc metalloproteinase
- [5] PMULT_1991 - unknown

Zusammenfassung

- Approximative Gencluster und ihre Anwendungen
- Software: Gecko 2

The screenshot displays the Gecko 2 software interface. At the top, there is a menu bar (File, View) and a toolbar with icons for file operations and animation. Below the toolbar is a search bar and a track view showing five tracks (1-5) with colored bars representing genomic data. Track 1 is highlighted in blue. Below the tracks is a table with columns: ID, #Genes, #Genomes, Score, and a list of coordinates. The table shows data for tracks 13 through 38. To the right of the table is a section titled 'Involved chromosomes' with a list of chromosome IDs (1-5) and their corresponding names. Below this is 'Global cluster information' showing 'Total distance: 0' and 'Score: 375,4019190375720'. Further down is 'Distance to center/median per dataset' with a row of buttons labeled 1, 2, 3, 4, 5. Below that is 'Genes in this Cluster' with a list of gene IDs and names, including HAEIN_1109, YFAST_0792, BAPHI_0207, ECOLI_0084, and PMULT_0130. The bottom of the window shows a 'Ready' status bar.

ID	#Genes	#Genomes	Score	
13	11	5	242.05	{322, 321, 320, 319, 861, 657, 101, 1273, 104, 546, 9, 307, 839, 545, 544, 838, 543, 836}
14	12	5	261.3	{321, 320, 322, 656, 319, 861, 657, 101, 1, 15, 11, 5
15	11	5	330.91	{322, 321, 320, 319, 861, 657, 101, 1273, 104, 546, 9, 307, 839, 545, 544, 838, 543, 836}
16	11	5	241.83	{322, 321, 320, 319, 861, 657, 101, 1273, 104, 546, 9, 307, 839, 545, 544, 838, 543, 836}
17	11	5	243.26	{322, 321, 320, 319, 861, 657, 101, 1273, 104, 546, 9, 307, 839, 545, 544, 838, 543, 836}
18	9	5	212.59	{322, 319, 861, 657, 321, 320, 101, 323, 3, 19, 10, 5
19	10	5	226.01	{320, 656, 655, 654, 319, 861, 657, 101, 1, 20, 11, 5
20	11	5	243.71	{321, 320, 656, 655, 654, 319, 861, 657, 1, 21, 9, 5
21	9	5	112.28	{320, 656, 655, 319, 861, 657, 101, 1273, 104, 546, 9, 307, 839, 545, 544, 838, 543, 836}
22	11	5	241.9	{322, 321, 320, 656, 655, 319, 861, 657, 1, 23, 10, 5
23	10	5	230.21	{321, 320, 656, 655, 319, 861, 657, 101, 1, 24, 9, 5
24	9	5	208.53	{322, 321, 320, 319, 861, 657, 101, 1273, 104, 546, 9, 307, 839, 545, 544, 838, 543, 836}
25	10	5	228.33	{322, 656, 321, 320, 319, 861, 657, 101, 3, 26, 9, 5
26	9	5	216.48	{656, 321, 320, 319, 861, 657, 101, 1273, 104, 546, 9, 307, 839, 545, 544, 838, 543, 836}
27	9	5	196.78	{322, 321, 320, 319, 861, 657, 101, 1273, 104, 546, 9, 307, 839, 545, 544, 838, 543, 836}
28	9	5	208.06	{322, 321, 320, 319, 861, 657, 101, 1273, 104, 546, 9, 307, 839, 545, 544, 838, 543, 836}
29	9	5	209.49	{322, 321, 320, 319, 861, 657, 101, 1273, 104, 546, 9, 307, 839, 545, 544, 838, 543, 836}
30	9	5	300.23	{307, 546, 104, 547, 9, 103, 1235, 548, 844, 31, 12, 5
31	12	5	358.48	{544, 545, 839, 307, 546, 104, 547, 9, 103, 32, 13, 5
32	13	5	375.4	{838, 544, 545, 839, 307, 546, 104, 547, 9, 33, 11, 5
33	11	5	341.31	{544, 545, 839, 307, 546, 104, 547, 9, 103, 34, 12, 5
34	12	5	358.49	{838, 544, 545, 839, 307, 546, 104, 547, 9, 35, 10, 5
35	10	5	309.98	{544, 545, 839, 307, 546, 104, 547, 9, 103, 36, 11, 5
36	11	5	327.56	{838, 544, 545, 839, 307, 546, 104, 547, 9, 37, 9, 5
37	9	5	296.81	{544, 545, 839, 307, 546, 104, 547, 9, 103, 38, 10, 5
38	10	5	314.61	{838, 544, 545, 839, 307, 546, 104, 547, 9, 103, 39, 10, 5}

Involved chromosomes

- 1 HAEIN
- 2 YFAST
- 3 BAPHI
- 4 ECOLI
- 5 PMULT

Global cluster information:

Total distance: 0
Score: 375,4019190375720

Distance to center/median per dataset:

1 0 2 3 4 5 0

Genes in this Cluster:

- 547 1 HAEIN_1109 - phospho-N-acetylauramoyl-pentapeptide- (HraY)
- 2 YFAST_0792 - phospho-N-acetylauramoyl-pentapeptide- trans
- 3 BAPHI_0207 - phospho-N-acetylauramoyl-pentapeptide- trans
- 4 ECOLI_0084 - phospho-N-acetylauramoyl-pentapeptide trans
- 5 PMULT_0130 - HraY
- 104 1 HAEIN_1111 - cell division protein (ftsM)
- 2 YFAST_0793 - cell division protein
- 3 BAPHI_0205 - cell division protein ftsM

Zusammenfassung

- Approximative Gencluster und ihre Anwendungen
- Software: Gecko 2

The screenshot displays the Gecko 2 software interface. The top part shows a heatmap with 5 rows and 20 columns of colored cells, representing gene clusters. Below the heatmap is a table with columns: ID, #Genes, #Genomes, and Score. The bottom right panel shows 'Global cluster information' for a selected cluster, including total distance, score, and a list of genes with their descriptions.

ID	#Genes	#Genomes	Score
0	9	5	262.56
1	9	5	243.83
2	9	5	219.19
3	9	5	221.12
4	11	5	289.16
5	19	5	452.4
6	20	5	486.71
7	22	5	508.1
8	28	5	607.51
9	9	5	240.74
10	11	5	290.42
11	17	5	406.87
12	9	5	235.71
13	11	5	242.05
14	12	5	261.3
15	11	5	230.91
16	11	5	241.83
17	11	5	243.26
18	9	5	212.59
19	10	5	226.01
20	11	5	243.71
21	9	5	212.28
22	11	5	241.9
23	10	5	230.21
24	9	5	208.53
25	10	5	228.33

Global cluster information:
Total distance: 7
Score: 262.552551249108

Distance to center/median per dataset:
1 2 3 4 5 1

Genes in this Cluster:

- 702 1 HAEIN_0467 - glucose-inhibited division protein (gidB)
- 4 ECOLI_3669 - phenotype: DNA - replication, repair, restr
- 5 PMULT_1486 - G1b
- 576 1 HAEIN_0464 - ATP synthase F0, subunit c (atpE)
- 2 YFAST_1145 - ATP synthase, C chain
- 3 BAPHI_0003 - ATP synthase C chain
- 4 ECOLI_3666 - enzyme: ATP-proton motive force interconvers
- 5 PMULT_1489 - AtpE
- 579 1 HAEIN_0462 - ATP synthase F1, subunit delta (atpH)
- 2 YFAST_1143 - ATP synthase, delta chain
- 3 BAPHI_0005 - ATP synthase delta chain
- 4 ECOLI_3664 - F1 sector of membrane-bound ATP synthase, de
- 5 PMULT_1491 - AtpH
- 377 1 HAEIN_0458 - ATP synthase F1, subunit epsilon (atpC)

- Verbesserung der praktischen Laufzeiten
- Hinzufügen der evolutionären Unterschiede innerhalb der Gencluster, um Consensus-Mengen zu verbessern
- Unabhängigkeit von Homologie-Zuweisungen
→ Sequenzähnlichkeiten der Gene statt feste Gruppierung
- Anwendung der Algorithmen auf andere Sequenztypen:
→ Sequenzen von Transkriptionsfaktor-Bindungsstellen
→ 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!