Genfamilienfreier Genomvergleich[1]

Daniel Dörr[2]

Abstract: Das Genom bezeichnet die gesamte genetische Information eines Organismus, welche hauptsächlich auf Chromosomen gespeichert ist. Der rechnergestützte Vergleich der Genome unterschiedlicher Spezies gewährt wertvolle Einsichten in deren gemeinsame und individuelle evolutionäre Historie. Hierzu werden Mutationen identifiziert, welche die DNA-Sequenzen in der evolutionären Vergangenheit verändert haben. Eine bestimmte Art von Mutationen führt zu Veränderungen in der Genreihenfolge von Genomen. Diese Arbeit stellt neue rechnergestützte Vergleichsmethoden der Genreihenfolge in Genomen unterschiedlicher Spezies vor. Hierzu werden klare Optimierungsprobleme identifiziert, deren Berechnungskomplexität analysiert und exakte, approximative, sowie heuristische Lösungsverfahren entwickelt.

1 Einleitung

Die *rechnergestützte vergleichende Genomik* gewährt wertvolle Einsichten in die gemeinsame und individuelle evolutionäre Historie von lebenden und ausgestorbenen Spezies. Genome zu vergleichen bedeutet, deren Unterschiede zu bestimmen, die durch Mutationen in ihrer evolutionären Vergangenheit entstanden sind.

Im Bereich der Genomevolution differenziert man zwischen *Punktmutationen*, *Genomumordnungen* und Änderungen des *Gengehalts* von Genomen. Punktmutationen verändern ein oder wenige aufeinanderfolgende Nukleotide in der DNA-Sequenz. Genomumordnungen ändern die Reihenfolge der Gene und ihre Aufteilung in chromosomale Sequenzen. Der Gengehalt wird durch die Evolution von Genfamilien beeinflusst, welche zu Genduplikationen oder dem Verlust von Genen führt.

Studien zur Erforschung von Genomumordnungen zwischen Genomen setzen die Kenntnis der evolutionären Verhältnisse zwischen deren Genen voraus. Mittels des biologischen Konzepts der *Homologie* kann die Menge aller Gene in Genfamilien unterteilt werden: Alle Gene in einer Genfamilie sind paarweise homolog zueinander, was bedeutet, dass sie von einer gemeinsamen Ursequenz abstammen. Homologien zwischen Genen sind in der Regel unbekannt und werden daher häufig mit rechnergestützten Methoden vorhergesagt. Dazu werden Sequenzähnlichkeiten zwischen Genen oder andere Ähnlichkeiten in den Eigenschaften ihrer Genprodukte quantifiziert. Allerdings ist die Vorhersage von Homologien häufig unzuverlässig, was zu Fehlern in einer anschließenden Studie von Genomumordnungen führt.

[1] Englischer Titel der Dissertation: “Gene Family-free Genome Comparison” [Do15]

[2] Universität Bielefeld, Technische Fakultät, Universitätsstr. 25, 33615 Bielefeld
aktuelle Adresse: EPFL, School of Computer and Communication Sciences, 1015 Lausanne, Schweiz

Diese Doktorarbeit verfolgt einen neuen Forschungszweig mit der Zielsetzung, Fehler durch falsche oder unvollständige Vorhersagen von Genfamilien in der Untersuchung von Genomumordnungen zu vermeiden. Dazu werden neue rechnergestützte Methoden zur Erforschung von Genomumordnungen entwickelt, die die Kenntnis von Genfamilien nicht voraussetzen. Dieser Ansatz, auch *genfamilienfreier Genomvergleich* genannt, greift direkt auf Genähnlichkeiten zurück, welche üblicherweise zur Vorhersage von Genfamilien verwendet werden. Somit können Unterschiede, welche durch Punktmutationen entstanden sind, in der Untersuchung von Genomumordnungen berücksichtigt werden.

2 Gene, Genome und Genähnlichkeiten

Die dem Genom eines Organismus zugehörigen DNA-Sequenzen beherbergen *vererbbare Eigenschaften*, welche messbare Funktionen im Zellsystem des Organismus ausüben und Gegenstand natürlicher Selektion sind. In dieser Arbeit wird ein Segment auf einer DNA-Sequenz, welches mit einer vererbbaren Eigenschaft assoziiert ist, *Gen* genannt. Zwei oder mehr Gene, die von derselben Ursequenz abstammen sind *homolog* [Fi00]. Eine *Genfamilie* bezeichnet eine Menge homologer Gene.

Im Folgenden ist ein Genom G gänzlich durch ein Tupel $G \equiv (\mathscr{C}, \mathscr{A})$ repräsentiert, wobei $\mathscr{C}$ eine nichtleere Menge eindeutiger Gene und $\mathscr{A}$ die Menge von (Gen-) *Nachbarschaften* sind. Gene werden repräsentiert durch ihre Extremitäten, d. h., ein Gen $g \equiv (g^{\mathrm{t}}, g^{\mathrm{h}})$, $g \in \mathscr{C}$, besteht aus einem *Ende (tail)* g^{t} und einem *Kopf (head)* g^{h}. Die Extremitäten zweier beliebiger Gene können jeweils eine *Nachbarschaft* formen und somit die Genmenge eines Genoms zu ein oder mehreren linearen oder zirkulären *Genreihenfolgen*, auch *Chromosomen* genannt, zusammensetzen. Die Endpunkte linearer Chromosomen, *Telomere* genannt, werden als spezielle "Gene" behandelt, welche nur eine Extremität "$\circ$" besitzen. Eine Genreihenfolge kann in zwei Richtungen gelesen werden, wobei eine der beiden Richtungen zur kanonischen *Leserichtung* bestimmt wird. Ein Gen g, welches dieser Leserichtung gegenläufig ist, d. h. dessen Kopf vor dem Ende erscheint, wird mit einem Balken $\overline{g}$ gekennzeichnet. Im Folgenden wird bequemer Weise von der Notation $\mathscr{C}(G)$ und $\mathscr{A}(G)$ Gebrauch gemacht, um Bezug auf die Mengen der Gene und Nachbarschaften eines Genomes G zu nehmen.

Das Prinzip des genfamilienfreien Genomvergleichs verkörpert die Idee, Genreihenfolgen zu analysieren, ohne vorher die Zugehörigkeit ihrer Gene zu Genfamilien zu kennen. Stattdessen wird auf *Genähnlichkeiten* zurückgegriffen, welche symmetrische und reflexive *Ähnlichkeitsmaße* $\sigma : \Sigma \times \Sigma \to \mathbb{R}_{\geq 0}$ über dem Universum aller Gene Σ darstellen [Br13]. Ganz unabhängig davon welches Maß zur Bestimmung der Ähnlichkeit zweier Gene benutzt wird, wird in dieser Arbeit eine hohe Ähnlichkeit als Indikator für Homologie betrachtet. Dabei dient der *Genähnlichkeitsgraph* als zentrale Datenstruktur:

Definition 1 (Genähnlichkeitsgraph [DTS12, Br13]) *Gegeben seien k Genome $G_1, \ldots, G_k$ und ein Ähnlichkeitsmaß σ. Dann ist der* Genähnlichkeitsgraph *ein gewichteter, ungerichteter k-partiter Graph $B = (V_1, V_2, \ldots, V_k, E)$, wobei jede Knotenmenge V_i, $1 \leq i \leq k$, die Genmenge des i-ten Genoms repräsentiert, d. h., $V_i = \mathscr{C}(G_i)$, und die Kantenmenge*

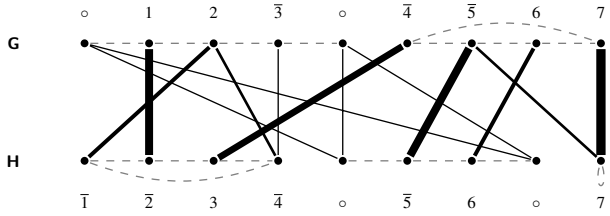

Abb. 1: Genähnlichkeitsgraph B zweier exemplarischer Genome G und H. Jeder Knoten repräsentiert ein Gen oder ein Telomer. Der Einfachheit halber sind Gene nach ihren Indices beschriftet. Schwarze Kanten kennzeichnen Ähnlichkeiten zwischen Genen aus G und H, wobei die Kantendicke den Ähnlichkeitsgrad visualisiert. Gestrichelte graue Kanten weisen auf Nachbarschaften in G und H hin, sind aber nicht Teil des Genähnlichkeitsgraphen.

$E = \{\{g,h\} \mid g \in \mathscr{C}(G_i),\ h \in \mathscr{C}(G_j), 1 \leq i < j \leq k : \sigma(g,h) > 0\}$ *Ähnlichkeiten zwischen Genen unterschiedlicher Genome darstellt. Dabei entspricht das Kantengewicht* $w(\{g,h\})$ *einer Kante* $\{g,h\} \in E$ *der Genähnlichkeit* $\sigma(g,h)$.

3 Genfamilienfreie Nachbarschaften

Die Anzahl *konservierter Nachbarschaften*, das heißt Nachbarschaften, welche zwei untersuchten Genomen gemein sind, kann als Maß zur Quantifizierung der Ähnlichkeit von Genomen verwendet werden. Gegeben seien zwei Genome G und H und Ähnlichkeitsmaß σ, zwei Nachbarschaften, $\{g_1^a, g_2^b\} \in \mathscr{A}(G)$ und $\{h_1^a, h_2^b\} \in \mathscr{A}(H)$ mit $a,b \in \{\mathrm{h}, \mathrm{t}\}$ sind *konserviert* wenn $\sigma(g_1,h_1) > 0$ und $\sigma(g_2,h_2) > 0$. Wenn der Gengehalt beider Genome identisch ist, dann ist die Anzahl konservierter Nachbarschaften das duale Maß zur *Breakpoint-Distanz* [Wa82]. Der *Score* der Nachbarschaft vier beliebiger Extremitäten g^a, h^b, i^c, j^d, wobei $a,b,c,d \in \{\mathrm{h}, \mathrm{t}\}$ und $g,h,i,j \in \Sigma$, ist als geometrisches Mittel ihrer entsprechenden Genähnlichkeiten definiert:

$$s(g^a, h^b, i^c, j^d) \equiv \sqrt{\sigma(g,h) \cdot \sigma(i,j)} \tag{1}$$

Ziel ist es, ein Matching im Genähnlichkeitsgraphen zweier Genome G und H zu etablieren, welches nicht nur Genähnlichkeiten berücksichtigt, sondern auch den Summenscore konservierter Nachbarschaften maximiert. Dabei wird zur Bestimmung der Gennachbarschaften die Matching-induzierte Genreihenfolge verwendet. Das bedeutet, dass Gene, deren entsprechende Knoten nicht Teil des Matchings sind, in der jeweiligen (ursprünglichen) Genreihenfolge übersprungen werden. Dieses Vorgehen erlaubt, duplizierte und neue Gene zu ignorieren und dadurch möglichst viele hochkonservierte Nachbarschaften zu identifizieren. Die Qualität des Matchings wird dabei mithilfe folgender Maße

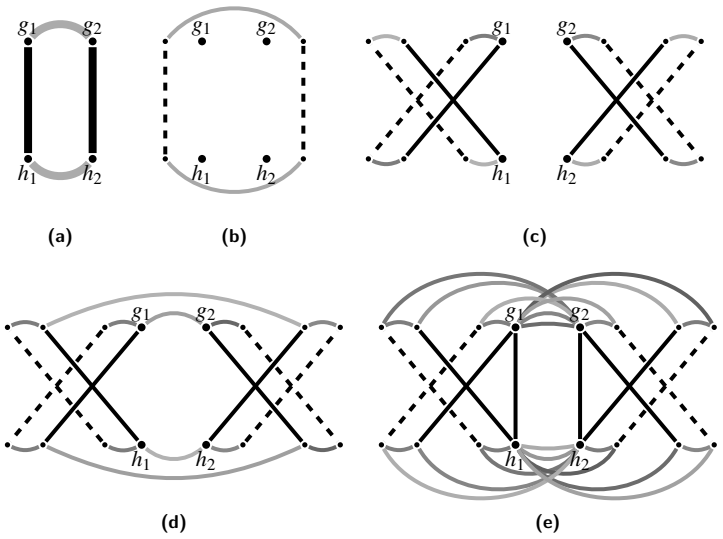

Abb. 2: Teile **(a)**–**(e)** zeigen alle möglichen Kandidaten konservierter Nachbarschaften der Gene g_1, g_2, h_1 und h_2. Schwarze Kanten sind dem Genähnlichkeitsgraphen B zugehörig, Paare von Bögen gleicher Farbe entsprechen konservierten Nachbarschaften der jeweiligen Gene. Gestrichelte schwarze Kanten weisen auf solche hin, deren Präsenz in B angenommen, jedoch nicht überprüft wird.

quantifiziert:

$$adj_{GH}(\mathscr{M}) = \sum_{\substack{\{\{g_1,h_1\},\{g_2,h_2\}\}\subseteq\mathscr{M},\\ \{g_1^a,g_2^b\}\in\mathscr{A}(G_{\mathscr{M}}),\\ \{h_1^a,h_2^b\}\in\mathscr{A}(H_{\mathscr{M}})}} s(g_1^a,g_2^b,h_1^a,h_2^b) \tag{2}$$

$$edg(\mathscr{M}) = \sum_{e\in\mathscr{M}} w(e) \tag{3}$$

Man beachte, dass die Genpaare (g_1,h_1) und (g_2,h_2) im Maß konservierter Nachbarschaften adj_{GH} die gleichen Extremitäten a und b haben, wobei $a,b\in\{\mathrm{h},\mathrm{t}\}$. Wir können nun folgendes Optimierungsproblem formulieren, welches zum Ziel hat, eine Lösung zu finden, die eine lineare Kombination beider erwähnter Qualitätsmaße adj_{GH} und edg maximiert:

Problem 1 (FF-Adjacencies) *Gegeben seien zwei Genome G, H und $\alpha\in[0,1]$, finde ein Matching $\mathscr{M}$ im Genähnlichkeitsgraphen B von G und H, sodass folgende Formel maximiert wird:*

$$\mathscr{F}_\alpha(\mathscr{M}) = \alpha\cdot adj_{GH}(\mathscr{M}) + (1-\alpha)\cdot edg(\mathscr{M}). \tag{4}$$

Theorem 1 *Problem FF-Adjacencies ist NP-schwer für* $0 < \alpha < \frac{1}{3}$.[3]

Mit dem ganzzahligen linearen Programm `FFAdj-2G` wurde ein exaktes Lösungsverfahren für Problem FF-Adjacencies entwickelt. Anhand einer umfassenden Analyse des Lösungsraums konnte die praktische Berechnungsgeschwindigkeit wesentlich verbessert werden, was in der Praxis den Vergleich bakterieller Genome ermöglicht. Dazu wurden unterschiedliche Ansätze gewählt, um optimale und strikt suboptimale Teilräume zu identifizieren. Als besonders effektiv hat sich hierbei die Suche nach einfachen, stark konservierten Nachbarschaften im Genähnlichkeitsgraphen zweier gegebener Genome herausgestellt. Abb. 2 skizziert hierbei die Vorgehensweise: Für jede stark konservierte Nachbarschaft zwischen vier Genen (Abb. 2 (a)), werden alle möglichen Kombinationen alternativer Lösungen identifiziert (Abb. 2 (b)–(e)), um eine Obergrenze für den Zugewinn im Summenscore konservierter Nachbarschaften zu berechnen. Liegt diese Obergrenze unterhalb des Scores der betrachteten Nachbarschaft, kann letztere als optimale Teillösung festgehalten werden. Solch identifizierte Nachbarschaften werden *Anker* genannt.

Programm `FFAdj-2G` und die ebenfalls entwickelte Heuristik `FFAdj-MCS` wurden auf simulierten und bakteriellen Genomdatensätzen getestet und mit einem genfamilienbasierten Lösungsverfahren verglichen [An09].

4 Genfamilienfreier Median

Im Folgenden wird die Problematik der Rekonstruktion von Ursequenzen im Rahmen des genfamilienfreien Genomvergleichs betrachtet. Die vorliegende Arbeit untersucht das Problem, ein viertes Genom M, *Median* genannt, anhand dreier gegebener Genome G, H, und I zu rekonstruieren. Hierbei wird das Modell des *gemischten multichromosomalen Breakpoint-Medians* verallgemeinert. Der Gengehalt des gesuchten Medians M ist wie folgt definiert: Jedes Gen $m \in \mathscr{C}(M)$ muss eindeutig mit einem Tripel von Genen (g,h,i), $g \in \mathscr{C}(G)$, $h \in \mathscr{C}(H)$ und $i \in \mathscr{C}(I)$ assoziiert sein. Des Weiteren verlangt die Berechnung der Scores konservierter Nachbarschaften Kenntnis der Genähnlichkeiten zwischen jedem Tripel von Genen (g,h,i) und dem jeweiligen, vermutlich ausgestorbenen, Gen m, wie in Abb. 3 (a) gezeigt wird. Da Genähnlichkeiten zu Mediangenen grundsätzlich nicht bekannt sind, werden sie von den jeweiligen Genen der gegebenen Genome abgeleitet. Hier folgen wir dem oben beschriebenen Score-Schema von Nachbarschaften und definieren die Ähnlichkeit zwischen Genen g, h und i zu ihrem entsprechenden Mediangen m als geometrisches Mittel ihrer paarweisen Genähnlichkeiten:

$$\sigma(g,m) = \sigma(h,m) = \sigma(i,m) \equiv \sqrt[3]{\sigma(g,h) \cdot \sigma(g,i) \cdot \sigma(h,i)} \tag{5}$$

Im Folgenden wird das Mapping $\pi_G(m) \equiv g$, $\pi_H(m) \equiv h$ und $\pi_I(m) \equiv i$ benutzt um von Mediangen m auf die entsprechenden Gene in den gegebenen Genomen Bezug zu nehmen. Zwei Mediangenkandidaten m_1 und m_2 sind *in Konflikt* wenn $m_1 \neq m_2$ und die Schnittmenge der assoziierten Gene $\{\pi_G(m_1), \pi_H(m_1), \pi_I(m_1)\}$ und $\{\pi_G(m_2), \pi_H(m_2), \pi_I(m_2)\}$ nichtleer ist. Folglich wird ein Median M *konfliktfrei* genannt, wenn keine zwei Gene $m_1, m_2 \subseteq \mathscr{C}(M)$ in Konflikt sind.

[3] In [Ko16] wurde die NP-schwere für allgemeines $\alpha > 0$ bewiesen.

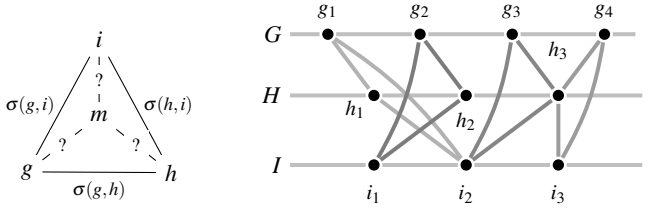

Abb. 3: **(a)** Visualisierung der Problematik der Berechnung von Genähnlichkeiten zu Mediangenen; **(b)** Beispiel dreier Genome G, H und I mit folgenden Mediangenkandidaten: $m_1 = (g_1, h_1, i_2)$ (gelb), $m_2 = (g_2, h_2, i_1)$ (rot), $m_3 = (g_3, h_3, i_2)$ (blau) und $m_4 = (g_4, h_3, i_3)$ (grün). Des Weiteren sind Mediangenkandidaten m_1, m_3 beziehungsweise m_3, m_4 in Konflikt.

Damit kann nun das Problem des genfamilienfreien Medians formalisiert werden:

Problem 2 (FF-Median) *Gegeben seien drei Genome G, H und I und ein Genähnlichkeitsmaß σ, finde einen konfliktfreien Median M, der folgende Formel maximiert:*

$$\mathscr{F}_{\lambda}(M) = \sum_{\{m_1^a, m_2^b\} \in \mathscr{A}(M)} \quad \sum_{\substack{X \in \{G,H,I\}, \\ \{\pi_X(m_1)^a, \pi_X(m_2)^b\} \in \mathscr{A}(X)}} s(m_1^a, m_2^b, \pi_X(m_1)^a, \pi_X(m_2)^b), \quad (6)$$

wobei $a, b \in \{h, t\}$ und $s(\cdot)$ der in Gleichung (1) definierte Nachbarschaftsscore ist.

Theorem 2 *Problem FF-Median ist NP-schwer.*

Zur exakten Lösung des Problems wurde das ganzzahlige lineare Program `FF-Median` entworfen. Weiterhin wurde das heuristische Verfahren `FFAdj-3G-H` entwickelt, welches zudem Änderungen der Genomsequenzen durch Genduplikation und -verlust toleriert. Diese zeigte im Vergleich auf simulierten Datensätzen überlegene Leistung. `FFAdj-3G-H` wurde anschließend zur Rekonstruktion von *Yersinia pestis* verwendet. Die Resultate wurden mit denen von Rajaraman *et al.* [RTC13] verglichen.

5 Genfamilienfreie Syntenie

Mit der Länge des evolutionären Zeitraums steigt die Zahl der Genomumordnungen, welche die Genreihenfolge zunehmend durchmischen. Aus diesem Grund sind Studien über evolutionär weit entfernte Genome, die konservierte Nachbarschaften identifizieren, nicht aufschlussreich. Dennoch können verallgemeinerte Definitionen konservierter Genreihenfolge ein schwächeres, aber dennoch vorhandenes Signal gemeinsamer Genreihenfolge auffangen. Dies ist Gegenstand eines Forschungszweigs, welcher sich mit der Identifikation *syntenischer Bereiche* beschäftigt. Wenn Genfamilien bekannt sind, dann lässt sich eine

Genreihenfolge als Zeichenfolge (String) über dem Alphabet von Genfamilienbezeichnungen darstellen. Ein Paar von Intervallen in zwei Strings wird *Common Intervals* genannt, wenn ihre Zeichenmenge identisch ist. Die Definition von Common Intervals wurde ursprünglich auf Permutationen eingeführt [UY00] und anschließend auf allgemeine Strings erweitert [Am03, Di07]. Common Intervals können zur Bestimmung syntenischer Bereiche in zwei oder mehr Genomen verwendet werden. Im Folgenden wird die Definition von Common Intervals auf *Indeterminate Strings* erweitert. Indeterminate Strings sind Sequenzen, in denen jede Position aus einer nicht-leeren Zeichenmenge besteht. Mehrere Modelle von Common Intervals für Indeterminate Strings werden vorgestellt und effiziente Algorithmen für das Auffinden entsprechender Intervallpaare in zwei Indeterminate Strings entwickelt. Diese neuen Algorithmen können zur Bestimmung syntenischer Bereiche im Rahmen des genfamilienfreien Genomvergleichs verwendet werden [Do14].

Für einen Indeterminate String S mit n Positionen muss gelten, dass für jedes i, $1 \leq i \leq n$, $S[i] \subseteq \Sigma$ und $S[i] \neq \emptyset$, wobei $S[i]$ die Zeichenmenge der i-ten Position in S ist. Im speziellen Fall, dass jede Position eines Indeterminate Strings S eine einelementige Menge ist, ist S äquivalent zu einem gewöhnlichen String. Die *Länge* eines Indeterminate Strings S mit n Positionen wird mit $|S| \equiv n$ angegeben und die Kardinalität, d. h. die Anzahl *aller* Zeichen in S, mit $\|S\| \equiv \sum_{i=1}^{n} |S[i]|$. Zwei Positionen a und b, $1 \leq a \leq b \leq |S|$, induzieren einen *Indeterminate Teilstring* $S[a,b] \equiv S[a]\,S[a+1]\,\ldots\,S[b]$.

Die Idee hinter Common Intervals ist der Vergleich von Strings, oder besser gesagt Teilstrings, auf Basis ihrer Zeichenmengen. Die Zeichenmenge eines gewöhnlichen Strings S ist definiert als $\mathscr{C}(S) \equiv \{S[i] \mid 1 \leq i \leq |S|\}$. Das äquivalente Konzept für Indeterminate Strings ist wie folgt definiert:

Definition 2 (Zeichenmenge) *Die* Zeichenmenge *eines Indeterminate Strings S der Länge n wird mit $\mathscr{C}(S) \equiv \bigcup_{i=1}^{n} S[i]$ angegeben.*

Man beachte, dass die Zeichenmenge $\mathscr{C}(S)$ und $\mathscr{C}(T)$ zweier Indeterminate Strings S und T identisch sein kann, jedoch sich keine zwei Positionen zwischen S und T die gleiche Zeichenmenge teilen. Das strikte Analogon für *Common Intervals* in Indeterminate String ist:

Definition 3 (Strikte Common Intervals) *Gegeben seien zwei Indeterminate Strings S und T, dann sind zwei Intervalle $[i,j]$ in S und $[k,l]$ in T* Strikte Common Intervals *wenn ihre Zeichenmengen $\mathscr{C}(S[i,j])$ und $\mathscr{C}(T[k,l])$ gleich sind.*

Eine abgeschwächte Definition, basierend auf der Schnittmenge von Zeichenmengen, ist wie folgt:

Definition 4 (Schwache Common Intervals) *Gegeben seien zwei Indeterminate Strings S und T, dann sind zwei Intervalle $[i,j]$ in S und $[k,l]$ in T* Schwache Common Intervals *mit gemeinsamer Zeichenmenge $C = \mathscr{C}(S[i,j]) \cap \mathscr{C}(T[k,l])$ wenn für jede Position x, $i \leq x \leq j$, gilt, dass $C \cap S[x] \neq \emptyset$, und für jede Position y, $k \leq y \leq l$, gilt, dass $C \cap T[y] \neq \emptyset$.*

Einem früheren Forschungszweig folgend, können strikte und schwache Common Intervals ferner erweitert werden, indem eine begrenzte Anzahl an abweichenden Positionen erlaubt wird:

Definition 5 (Approximativ-schwache Common Intervals) *Gegeben seien zwei Indeterminate Strings S und T und Schwellenwert $\delta \in \mathbb{N}_0$. Dann sind zwei Intervalle, $[i,j]$ in S und $[k,l]$ in T, sind* approximativ-schwache Common Intervals *mit gemeinsamer Zeichenmenge $C = \mathscr{C}(S[i,j]) \cap \mathscr{C}(T[k,l])$ wenn die Anzahl der Positionen mit leerer Schnittmenge zu C durch δ begrenzt ist, d. h., $|\{x \mid i \leq x \leq j : S[x] \cap C = \emptyset\}| + |\{y \mid k \leq y \leq l : T[y] \cap C = \emptyset\}| \leq \delta$.*

Grundsätzlich listen Algorithmen zum Auffinden von Common Intervals in gewöhnlichen Strings nur solche Intervallpaare, die auch *maximal sind*, d. h., die nicht links oder rechts erweitert werden können, ohne die gemeinsame Zeichenmenge zu vergrößern. Die äquivalente Bedingung für Indeterminate Strings ist wie folgt:

Definition 6 (Maximal) *Ein Intervall $[i,j]$ in S ist* maximal *wenn (i) $i = 1$ oder $S[i-1] \not\subseteq \mathscr{C}(S[i,j])$ und (ii) $j = |S|$ oder $S[j+1] \not\subseteq \mathscr{C}(S[i,j])$.*

Man beachte, dass die Maximalitätseigenschaft nicht mit schwachen Common Intervals kombiniert werden kann, ohne dass sinnvolle Intervallpaare dabei verloren gehen. Stattdessen könnte man überlegen, den Suchraum über zwei Indeterminate Strings auf solche Paare zu begrenzen, die nicht im Bezug auf ihre gemeinsame Zeichenmenge erweitert werden können. Dies führt zu folgender Eigenschaft, welche von [Ja11] abgeleitet wurde:

Definition 7 (C-abgeschlossen) *Gegeben sei ein Indeterminate String S, ein Intervall $[i,j]$ und eine Zeichenmenge $C \subseteq \Sigma$. Dann ist das Intervall $[i,j]$ C-*abgeschlossen*, wenn $S[i]$, $S[j] \cap C \neq \emptyset$ und wenn $i = 1$ oder $S[i-1] \cap C = \emptyset$ und wenn $j = |S|$ oder $S[j+1] \cap C = \emptyset$.*

Allerdings ist die Anzahl der Intervallpaare, die mit Bezug auf ihre gemeinsame Zeichenmenge abgeschlossen sind, immer noch absurd hoch, auch für einfache schwache Indeterminate Strings. Eine vernünftige Balance zwischen dem Ausschluss unnötiger und dem Einschluss sinnvoller schwacher Common Intervals findet man in der Teilmenge solcher, die *beidseitig-abgeschlossen* sind:

Definition 8 (beidseitig-abgeschlossen) *Gegeben seien zwei Indeterminate Strings S und T und zwei Intervalle, $[i,j]$ in S und $[k,l]$ in T. Dann ist das Intervallpaar $[i,j],[k,l]$* beidseitig-abgeschlossen *wenn Intervall $[i,j]$ $\mathscr{C}(T[k,l])$-abgeschlossen und Intervall $[k,l]$ $\mathscr{C}(S[i,j])$-abgeschlossen ist.*

Folglich beschränken wir die Aufzählung schwacher Common Intervals und approximativ-schwacher Common Intervals auf solche, die beidseitig-geschlossen sind. Für alle erwähnten Varianten von Common Interals in Indeterminate Strings wurden Algorithmen entwickelt, deren Laufzeiten und Speicherverbrauch in folgenden drei Theoremen festgehalten ist:

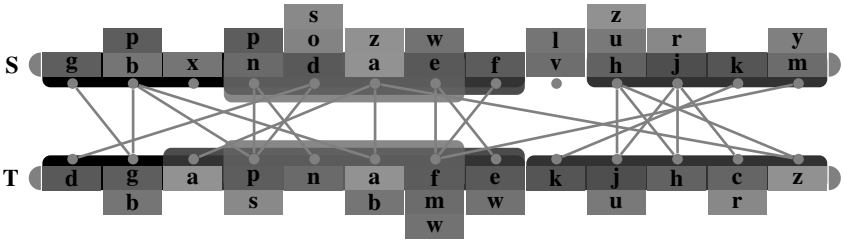

Abb. 4: Beispiel zweier Indeterminate Strings S und T. Folgende Intervallpaare sind ohne Anspruch auf Vollständigkeit hervorgehoben: Rote Intervallpaare kennzeichnen schwache, Blaue beidseitig-geschlossene schwache, und Schwarze beidseitig-geschlossene approximativ-schwache Common Intervals.

Theorem 3 *Alle Paare maximal strikter Common Intervals in zwei Indeterminate Strings S und T können in $O(|S| \cdot (\|S\| + \|T\|))$ Zeit und $O(\|S\| + \|T\|)$ Platz aufgezählt werden.*

Theorem 4 *Alle Paare beidseitig-geschlossener schwacher Common Intervals in zwei Indeterminate Strings S und T können in $O(|S|^2 \cdot |T|)$ Zeit und $O(|S| \cdot |T|)$ Platz aufgezählt werden.*

Theorem 5 *Gegeben sei ein Schwellenwert $\delta \geq 0$, dann können alle Paare beidseitig-geschlossener approximativ-schwacher Common Intervals in zwei Indeterminate Strings S und T in $O((\delta+1)^2 \cdot |S|^3 \cdot |T|)$ Zeit und $O((\delta+1)^2 \cdot |S| \cdot |T|)$ Platz aufgezählt werden.*

Die Algorithmen zum Auffinden schwacher bzw. approximativ-schwacher Common Intervals wurden implementiert und zur Analyse von *Genclustern* in bakteriellen Genomen verwendet. Gencluster sind kurze konservierte Bereiche funktional zusammenhängender Gene.

6 Zusammenfassung und Ausblick

Mit dieser Arbeit wurde ein neuer Forschungszweig der rechnergestützten vergleichenden Genomik angestoßen, dessen Ziel die Entwicklung neuer Vergleichsmethoden der Genreihenfolge von Genomen ist, die jedoch keine Kenntnis von Genfamilien voraussetzen. Dabei wurden Modelle und Lösungsverfahren für drei verschiedene Problemstellungen ausgearbeitet.

Im praktischen Teil der vorliegenden Arbeit wurden Genähnlichkeiten mittels einer Heuristik für lokales Sequenzalignment berechnet. Alternativ könnten hier Substitutionsratenfunktionen von Modellen der DNA-Evolution verwendet werden. Diese würden es ermöglichen, Genomumordnungsmodelle mit Modellen der DNA-Evolution zu kombinieren und würden somit dem Ziel einer ganzheitlichen Studie der Genomevolution ein Stück näherkommen.

Literaturverzeichnis

[Am03] Amir, A; Apostolico, A; Landau, G M; Satta, G: Efficient text fingerprinting via Parikh mapping. J. Discr. Alg., 1(5–6):409–421, 2003.

[An09] Angibaud, S; Fertin, G; Rusu, I; Thévenin, A; Vialette, S: On the Approximability of Comparing Genomes with Duplicates. J. Graph Alg. Appl., 13(1):19–53, 2009.

[Br13] Braga, M D V; Chauve, C; Doerr, D; Jahn, K; Stoye, J; Thévenin, A; Wittler, R: The Potential of Family-Free Genome Comparison. In: Models and Algorithms for Genome Evolution, Jgg. 19 in Comp. Biol., Kapitel 13, S. 287–323. Springer London, 2013.

[Di07] Didier, G; Schmidt, T; Stoye, J; Tsur, D: Character sets of strings. J. Discr. Alg., 5(2):330–340, 2007.

[Do14] Doerr, D; Stoye, J; Böcker, S; Jahn, K: Identifying gene clusters by discovering common intervals in indeterminate strings. BMC Genomics, 15(Suppl 6):S2, 2014.

[Do15] Doerr, D.: Gene Family-free Genome Comparison. Ph. D. thesis, Faculty of Technology, Bielefeld University, Germany, 2015.

[DTS12] Doerr, D; Thévenin, A; Stoye, J: Gene family assignment-free comparative genomics. BMC Bioinformatics, 13(Suppl 19):S3, 2012.

[Fi00] Fitch, Walter M: Homology a personal view on some of the problems. Trends Genet., 16(5):227–231, 2000.

[Ja11] Jahn, K: Efficient Computation of Approximate Gene Clusters Based on Reference Occurrences. J. Comput. Biol., 18(9):1255–1274, 2011.

[Ko16] Kowada, L A B; Doerr, D; Dantas, S; Stoye, J: New Genome Similarity Measures based on Conserved Gene Adjacencies. In: Proc. of RECOMB 2016, to appear. LNBI, Springer Verlag, Berlin, 2016.

[RTC13] Rajaraman, A; Tannier, E; Chauve, C: FPSAC: Fast phylogenetic scaffolding of ancient contigs. Bioinformatics, 29(23):2987–2994, 2013.

[UY00] Uno, T; Yagiura, M: Fast algorithms to enumerate all common intervals of two permutations. Algorithmica, 26(2):290–309, 2000.

[Wa82] Watterson, GA; Ewens, WJ; Hall, TE; Morgan, A: The Chromosome Inversion Problem. J. Theor. Biol., 99(1):1–7, 1982.

Daniel Dörr, geboren am 1. Juni 1983 in Saarbrücken, absolvierte sein Bachelor- und Masterstudium in Bioinformatik und Genomforschung an der Universität Bielefeld. Während des Masterstudiums war er Gaststudent am Technion in Haifa, Israel. Daniel war Doktorand des *CLIB-Graduiertencluster "Industrielle Biotechnologie"* und assoziierter Student des GRKs *Computational Methods for the Analysis of the Diversity and Dynamics of Genomes*. Während seiner Promotion machte er Forschungsaufenthalte bei IBM Research in Almaden, USA, und an der Simon Fraser University in Burnaby, Kanada. Letzterer wurde durch das DAAD "FITweltweit" Programm ermöglicht. Seit Mai 2015 ist Daniel Postdoctoral Fellow an der EFPL in Lausanne, Schweiz.