

Kapitel 7: Sequenzen-Alignment in der Bioinformatik

7.5: Maximum Weight Trace

VO Algorithm Engineering
 Professor Dr. Petra Mutzel
 Lehrstuhl für Algorithm Engineering, LS11

21. VO 21. Juni 2007

Literatur für diese VO

- A. Bockmayr, K. Reinert: Mathematische Aspekte der Bioinformatik, Skript WS 2005, Kap. 2: Alignments using Combinatorial Optimization, FU Berlin
- K. Reinert, H.-P. Lenhof, P. Mutzel, K. Mehlhorn, J.D. Kececioğlu: A Branch-and-Cut Algorithm for Multiple Sequence Alignment, RECOMB 1997, Santa Fe, New Mexico
- (E. Althaus, A. Caprara, H.-P. Lenhof, K. Reinert: A Branch-and-Cut Algorithm for Multiple Sequence Alignment, Mathematical Programming 2005)

Überblick

Einführung: Maximum Weight Trace

Graphentheoretische Formulierung

Formulierung als ILP

Schnittebenenverfahren

Branch-and-Cut Algorithmus

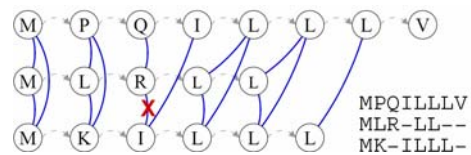
Experimentelle Ergebnisse

7.5: Maximum Weight Trace

Ziel: multiples Alignment, das am besten zu allen optimalen paarweisen Alignments passt

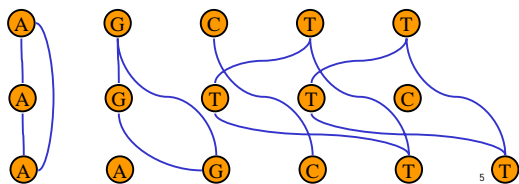
- Formulierung als Maximum Weight Trace Problem (Kececioğlu, 1993): *Trace*, das die Summe des Gewichts der realisierten Kanten maximiert

Alignmentgraph für 3 Sequenzen



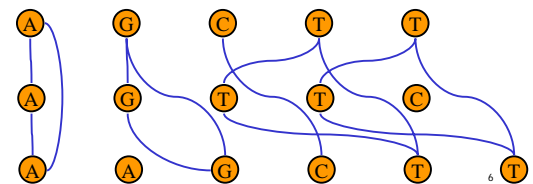
Maximum Weight Trace: Begriffe

- Def: **Vollständiger Alignment Graph (VAG):** vollständiger k-partiter Graph $G=(V,E)$ mit Kantenkosten $w(e) = p(i,j)$ (Wert einer Alignment von s_i mit t_j)
- Def: **Alignment Graph (AG):** Teilgraph des VAG
- Eine Kante (i,j) eines AG heißt **realisiert** bzgl. einer Alignment A , falls in A das Zeichen s_i mit t_j aligniert ist.



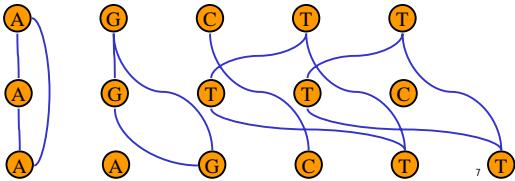
Maximum Weight Trace: Begriffe

- Eine Kante (i,j) eines AG heißt **realisiert** bzgl. einer Alignment A , falls in A das Zeichen s_i mit t_j aligniert ist.
- Die Menge aller realisierter Kanten heißt **Trace** der Alignment.
- Eine Kantenmenge $T \subseteq E$ heißt **Trace**, wenn eine Alignment existiert, die T realisiert.



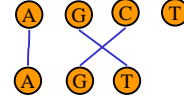
Maximum Weight Trace: Problem

- Geg: k Sequenzen und der dazugehörige Alignment Graph $G=(V,E)$ mit Kantenkosten w .
- Das Maximum Weight Trace Problem sucht einen Trace mit maximalem Gewicht $c(T) = \sum_{e \in T} w(e)$



Charakterisierung von Traces

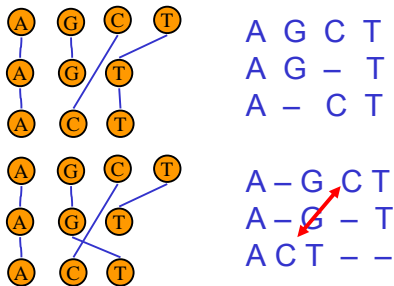
- Ist jeder Teilgraph von AG ein Trace eines Alignments?
- Nein! Z.B.:



- Ziel: Charakterisiere die Menge der zulässigen Traces:
- Ein Teilgraph T eines AG ist ein Trace genau dann wenn ...
- Für $k=2$ Sequenzen:
- ... wenn T keine Kreuzungen enthält (Einbettung wie von Sequenzen vorgegeben, geradlinige Kanten).

Charakterisierung von Traces

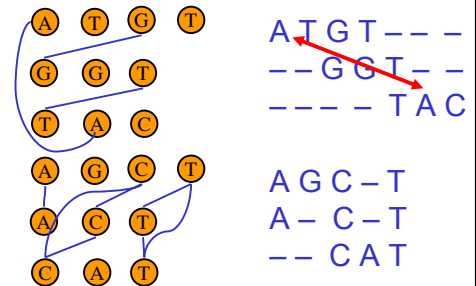
- Ein Teilgraph T eines AG ist ein Trace genau dann wenn ...



9

Charakterisierung von Traces

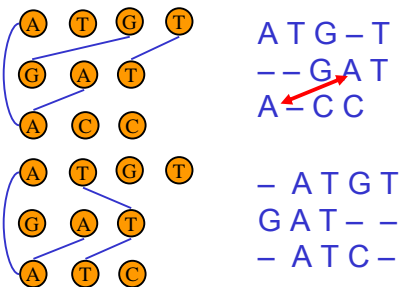
- Ein Teilgraph T eines AG ist ein Trace genau dann wenn ...



10

Charakterisierung von Traces

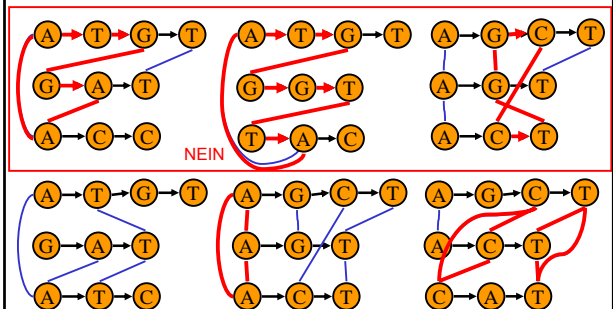
- Ein Teilgraph T eines AG ist ein Trace genau dann wenn ...



11

Charakterisierung von Traces

- Ein Teilgraph T eines AG ist ein Trace genau dann wenn ...

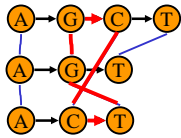


- ... T keinen **mixed cycle** (enthält ≥ 1 Kante aus H) in EAG hat

12

Charakterisierung von Traces

- Idee: Erweiterung des AG durch Kantenmenge H → **Erweiterter Alignment Graph (EAG)**
- Ein Kreis C heißt **mixed**, wenn er mindestens eine Kante aus H enthält.
- Ein Teilgraph T eines AG ist ein Trace genau dann wenn T keinen **mixed cycle** in EAG enthält.



A - G C T
 A - G - T
 A C T - -

13

Formulierung des MWT als ILP

$$\begin{aligned} \max \quad & \sum_{e \in E} w_e x_e \\ \sum_{e \in C \cap E} x_e & \leq |E \cap C| - 1 \quad \forall \text{ mixed cycles } C \text{ in EAG} \\ 0 & \leq x_e \leq 1 \quad \forall e \in E \\ x_e & \text{ ganzzahlig} \end{aligned}$$

Die nicht-trivialen Ungleichungen heißen **Mixed Cycle Ungleichungen**.

15

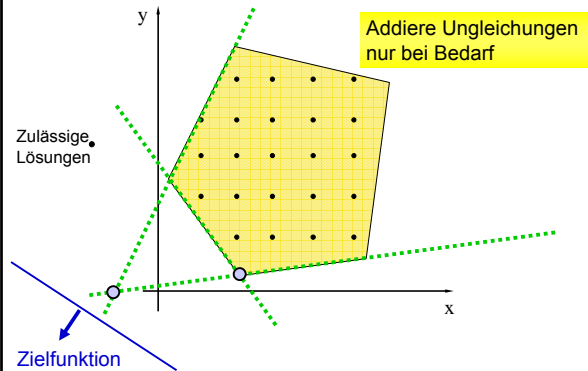
Idee von Schnittebenenverfahren

- (1) Starte mit einer Teilmenge der Restriktionen
- (2) Löse LP, sei x^* die gefundene Optimallösung
- (3) Entscheide, ob es weggelassene Restriktionen $a^T x \leq b_0$ gibt, so dass $a^T x > b_0$?
 - (3.1) Falls NEIN: STOP (Relaxierung gelöst)
 - (3.2) Falls JA: Bestimme solche, füge sie zu LP hinzu und gehe zu (1)

Separationsproblem

16

Schnittebenenverfahren

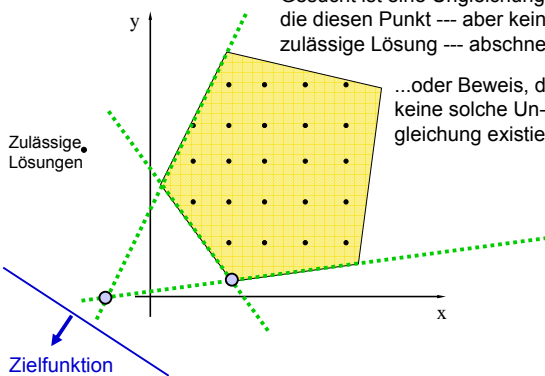


17

Separationsproblem

Gegeben ist ein Punkt x und OP. Gesucht ist eine Ungleichung, die diesen Punkt --- aber keine zulässige Lösung --- abschneidet...

...oder Beweis, dass keine solche Ungleichung existiert.



18

Satz von Grötschel, Lovasz, Schrijver

Das Optimierungsproblem ist in polynomieller Zeit lösbar genau dann wenn das zugehörige Separationsproblem in polynomieller Zeit lösbar ist.

Frage: Können wir das Separationsproblem für Mixed Cycle Ungleichungen in polynomieller Zeit lösen?

$$\sum_{e \in C \cap E} x_e \leq |E \cap C| - 1 \quad \forall \text{ mixed cycles } C \text{ in EAG}$$

Ja!

19

Definition Separationsproblem

Gegeben ist ein Punkt $\bar{x} \in \mathbb{R}^E$ und ein Polytop P .
Bestimme, ob $\bar{x} \in P$, und falls $\bar{x} \notin P$, finde eine Restriktion $a^T x \leq b_0$, die für alle Punkte $x \in P$ gültig ist, jedoch für den Punkt \bar{x} verletzt ist.

D.h. für alle Punkte $x \in P$ muß gelten $a^T x \leq b_0$ und für \bar{x} muß gelten $a^T \bar{x} > b_0$

20

Separierung der Mixed Cycle Ungleichungen

- Geg. ist eine Lösung x^* der LP-Relaxierung. Finde einen mixed cycle, der die Ungleichung verletzt oder beweise, dass kein solcher existiert.

$$\sum_{e \in C \cap E} x_e \leq |E \cap C| - 1 \iff \sum_{e \in C \cap E} (1 - x_e) \geq 1$$

- Jede Kante $e \in E$ erhält Kosten $c_e = 1 - x_e^*$ und die gerichteten Kanten $a \in H$ erhalten $c_a = 0$.
- Dann berechnen wir für jeden Knoten s_{ij} (entspricht j-tes Zeichen in Sequenz i) einen kürzesten Weg von s_{ij+1} nach s_{ij} bzgl. den Kosten c .
- Falls die Kosten von P kleiner als 1 sind, dann haben wir eine verletzte Ungleichung gefunden: der Kreis $P \cup (s_{ij}, s_{ij+1})$. Ansonsten existiert kein solcher Kreis durch diese H-Kante.

Das Trace-Polytop

$$P_T(G) = \text{conv}\{\chi^T \in \{0, 1\}^E \mid T \subseteq E \text{ ist ein Trace in } G\}$$

Ziel: Finde Maximum Weight Trace:

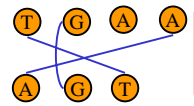
$$\max\{w^T x \mid x \in P_T(G)\}$$

- Aufgabe: Finde eine partielle Beschreibung durch Ungleichungen des Trace Polytops
- Lösung:
 - Mixed Cycle Ungleichungen (aus ILP-Beschreibung)
 - Cliquen-Ungleichungen

22

Verstärken der LP-Relaxierung durch zusätzliche Ungleichungen

- Für diese Ungleichungen betrachten wir jeweils 2 Sequenzen
- Definiere den Graphen G'' : Knoten entsprechen den E-Kanten, und zwei Knoten sind durch eine Kante miteinander verbunden, wenn sie sich kreuzen \rightarrow **Cliquen-Ungleichungen**:



$$\sum_{e \in K} x_e \leq 1 \quad \text{für alle Cliquen } K \text{ in } G''$$

Ja! Mithilfe von Pairgraphen

Frage: Können wir das Separationsproblem für Cliquen-Ungleichungen in polynomieller Zeit lösen?

23

MWT für k=2 Sequenzen

- Für k=2 Sequenzen: Die LP-Relaxierung gegeben durch die trivialen Ungleichungen und die Cliquen-Ungleichungen ist ganzzahlig und liefert als Ergebnis einen optimalen Trace.

$$\begin{aligned} \max \quad & \sum_{e \in E} w_e x_e \\ \text{s.t.} \quad & \sum_{e \in K} x_e \leq 1 \quad \forall \text{ Cliquen } K \text{ in } G'' \\ & 0 \leq x_e \leq 1 \quad \forall e \in E \end{aligned}$$

~~x_e ganzzahlig~~

- Dies liefert somit einen polynomiellen Algorithmus, der nicht auf Dynamischer Programmierung beruht. (Lange Zeit offenes Problem).

24

Das Trace-Polytop

$$P_T(G) = \text{conv}\{\chi^T \in \{0, 1\}^E \mid T \subseteq E \text{ ist ein Trace in } G\}$$

Ziel: Finde Maximum Weight Trace:

$$\max\{w^T x \mid x \in P_T(G)\}$$

- Aufgabe: Finde eine partielle Beschreibung durch Ungleichungen des Trace Polytops
- Lösung:
 - Mixed Cycle Ungleichungen (aus ILP-Beschreibung)
 - Cliquen-Ungleichungen
 - Ladder-Ungleichungen, ...

25

Branch-and-Cut Verfahren

Verbindung von Schnittebenenverfahren mit Branch-and-Bound

Versuche, jeweils die Teilprobleme (LP-Relaxierungen) mittels Schnittebenenverfahren zu lösen

Falls die Lösung nicht ganzzahlig ist, dann wähle nicht-ganzzahlige Variable und generiere zwei neue Teilprobleme:

P1 mit zusätzlichen Restriktionen $x_c=0$
 P2 mit zusätzlichen Restriktionen $x_c=1$

26

Multiple Sequence Alignment mit beliebigen Gap-Kosten

- Das ILP-Modell kann erweitert werden um beliebige affine Gap-Funktionen. Hierzu werden zusätzliche Bögen in EAG eingefügt, die die Gaps modellieren.
- u.a. auch Affine Funktion: $w(k) = h + gk$, mit $w(0) = 0$
- auch beliebige andere Gap-Funktionen
- auch ortsabhängig



27

Branch-and-Cut Algorithmus für MWT

- Preprocessing zur Variablenreduktion
- LP-basierte primale Heuristik
- Separationsverfahren
- verschiedene Branching- und Enumerationsregeln

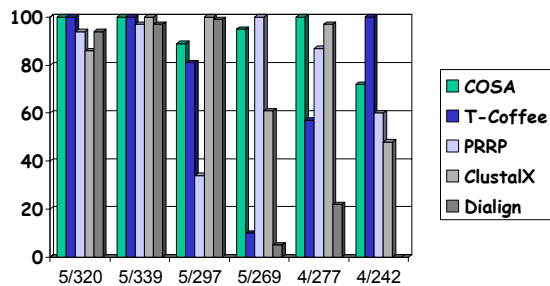
28

Experimentelle Resultate

- COSA (**C**ombinatorial **S**equencing **A**lignment): Branch-and-Cut Algorithmus mit beliebigen Gap-Funktionen
- beweisbar optimal gelöst werden konnten 18 aus 27
- auch die anderen sehr nahe am Optimum
- größte gelöste Instanz:
 - 5 Sequenzen mit bis zu 572 Zeichen insgesamt
 - ohne Gap-Kosten 1997: 15 Sequenzen der Länge ca. 4000, 6 Sequenzen mit ca. 1800
- Vergleich (als Heuristik) mit den besten Aligrierungsprogrammen: PRRP, ClustalX, Dialign, T-Coffee (Heuristiken, default Parameter)
- Problem: LPs werden zu groß!

29

Experimentelle Resultate



Y-Achse: Prozente relativ zum besten Programm

Ende