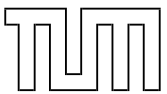
Algorithmische Bioinformatik 1

Dr. Hanjo Täubig

Lehrstuhl für Effiziente Algorithmen
(Prof. Dr. Ernst W. Mayr)
Institut für Informatik
Technische Universität München

Sommersemester 2009



Übersicht

1. Paarweises Sequenzen-Alignment
   - Distanz- und Ähnlichkeitsmaße

Restriktion

Definition (Restriktion)

Sei $u \in \overline{\Sigma}^*$. Dann sei die Restriktion von u auf Σ mit Hilfe eines Homomorphismus h wie folgt definiert

$$u|_\Sigma = h(u), \text{ wobei}$$

$$\begin{aligned} h(a) &= a && \text{für alle } a \in \Sigma, \\ h(-) &= \epsilon, && \\ h(u'u'') &= h(u')h(u'') && \text{für alle } u', u'' \in \Sigma^*. \end{aligned}$$

Die Restriktion von $u \in \overline{\Sigma}^*$ auf Σ ist also nichts anderes als das Löschen aller Leerzeichen $(-)$ aus u.

Alignment

Definition (Alignment)

Ein (paarweises) Alignment ist ein Paar $(\overline{a}, \overline{b}) \in \overline{\Sigma}^* \times \overline{\Sigma}^*$ mit

- $|\overline{a}| = |\overline{b}|$ und
- $\overline{a}_i \neq - \neq \overline{b}_i$ für alle $i \in [1 : |\overline{a}|]$ mit $a_i = b_i$.

$(\overline{a}, \overline{b})$ ist ein Alignment für $a, b \in \Sigma^*$, wenn $\overline{a}|_\Sigma = a$ und $\overline{b}|_\Sigma = b$.

Beispiel

$$\begin{array}{cccccccc} A & - & G & G & C & A & T & T \\ A & G & C & G & C & - & T & T \end{array}$$

Alignment von $AGGCATT$ mit $AGCGCTT$
Dieses Alignment hat Distanz 3,
es gibt jedoch ein besseres mit Distanz 2.

Alignment-Kosten und Alignment-Distanz

Definition

Sei $\overline{w} : \overline{\Sigma}_0^2 \to \mathbb{R}_+$ eine Kostenfunktion (für einzelne Zeichen).

Die Notation von $\overline{w}$ wird wie folgt auf Sequenzen erweitert, um die (Gesamt-)Kosten eines Alignments $(\overline{a}, \overline{b})$ für (a, b) zu definieren:

$$\overline{w}(\overline{a}, \overline{b}) = \sum_{i=1}^{|\overline{a}|} \overline{w}(\overline{a}_i, \overline{b}_i).$$

Die Alignment-Distanz von $a, b \in \Sigma^*$ ist definiert als

$$\overline{d}_{\overline{w}}(a, b) := \min \left\{ \overline{w}(\overline{a}, \overline{b}) \ : \ (\overline{a}, \overline{b}) \text{ ist Alignment für } a, b \right\}.$$

Beziehung zwischen Edit- und Alignment-Distanz

Lemma

Sei $w : \overline{\Sigma}_0^2 \to \mathbb{R}_+$ *eine Kostenfunktion (die hier sowohl für die Edit-Operationen als auch für das Alignment benutzt wird, also* $w = \overline{w}$*) und seien* $a, b \in \Sigma^*$.

Für jedes Alignment $(\overline{a}, \overline{b})$ *von* a *und* b *gibt es eine Folge* s *von Edit-Operationen, so dass* $a \overset{s}{\Rightarrow} b$ *und* $w(s) = w(\overline{a}, \overline{b})$

Beziehung zwischen Edit- und Alignment-Distanz

Beweis.

Sei $(\overline{a}, \overline{b})$ ein Alignment für a und b.
Betrachte die Folge $s = (s_1, \ldots, s_{|\overline{a}|})$ von Edit-Operationen mit $s_i = (\overline{a}_i, \overline{b}_i)$.
Offensichtlich gilt: $a \overset{s}{\Rightarrow} b$. Für die Edit-Kosten erhalten wir somit:

$$w(s) = \sum_{i=1}^{|\overline{a}|} w(s_i) = \sum_{i=1}^{|\overline{a}|} w(\overline{a}_i, \overline{b}_i) = w(\overline{a}, \overline{b})$$

□

Beziehung zwischen Edit- und Alignment-Distanz

Aus diesem Lemma folgt sofort, dass die Edit-Distanz von zwei Zeichenreihen höchstens so groß ist wie die Alignment-Distanz.

Folgerung

Sei $w : \overline{\Sigma}_0^2 \to \mathbb{R}_+$ *eine Kostenfunktion, dann gilt für alle* $a, b \in \Sigma^*$*:*

$$d_w(a, b) \leq \overline{d}_w(a, b)$$

Beziehung zwischen Edit- und Alignment-Distanz

Lemma

Sei $w : \overline{\Sigma}_0^2 \to \mathbb{R}_+$ *eine metrische Kostenfunktion und seien* $a, b \in \Sigma^*$.

Für jede Folge von Edit-Operationen mit $a \overset{s}{\Rightarrow} b$ *gibt es ein Alignment* $(\overline{a}, \overline{b})$ *von* a *und* b, *so dass* $w(\overline{a}, \overline{b}) \leq w(s)$.

Beziehung zwischen Edit- und Alignment-Distanz

Beweis.

(durch Induktion über $n = |s|$)

Induktionsanfang ($n = 0$):

- Aus $|s| = 0$ folgt, dass $s = \epsilon$.
- Also ist $a = b$ und $w(s) = 0$.
- Wir setzen nun $\overline{a} = a = b = \overline{b}$ und erhalten ein Alignment $(\overline{a}, \overline{b})$ für a und b mit $w(\overline{a}, \overline{b}) = 0 \leq w(s)$.

Beziehung zwischen Edit- und Alignment-Distanz

Beweis.

Induktionsschritt ($n \longrightarrow n+1$):

- Sei $s = (s_1, \ldots, s_n, s_{n+1})$ eine Folge von Edit-Operationen mit $a \overset{s}{\Rightarrow} b$.
- Sei nun $s' = (s_1, \ldots, s_n)$ und $a \overset{s'}{\Rightarrow} c \xrightarrow{s_{n+1}} b$ für ein $c \in \Sigma^*$.
- Aus der Induktionsvoraussetzung folgt nun, dass es ein Alignment $(\overline{a}, \overline{c})$ von a, c gibt, so dass $w(\overline{a}, \overline{c}) \leq w(s')$.

Beziehung zwischen Edit- und Alignment-Distanz

Beweis.

- Betrachte zuerst den Fall, dass die letzte Edit-Operation $s_{n+1} = (x, y)$ eine Substitution, ein Match oder eine Deletion ist, d.h. $x \in \Sigma$ und $y \in \overline{\Sigma}$.
- Wir können dann (wie in der Abbildung) ein Alignment für a und b erzeugen, indem wir die Zeichenreihe $\overline{b}$ geeignet aus $\overline{c}$ unter Verwendung der Edit-Operation(x, y) umformen.

$\overline{a}$		$*$	
$\overline{c}$		x	
$\overline{b}$		y	

$\overline{a}|_\Sigma = a$
$\overline{c}|_\Sigma = c$
$\overline{b}|_\Sigma = b$

Skizze: $s_{n+1} = (x, y)$ ist eine Substitution, Match oder Deletion

Beziehung zwischen Edit- und Alignment-Distanz

Beweis.

Es gilt dann:

$$
\begin{aligned}
w(\overline{a},\overline{b}) &= w(\overline{a},\overline{c}) \underbrace{-w(\overline{a}_i,\overline{c}_i)+w(\overline{a}_i,\overline{b}_i)}_{\leq w(\overline{c}_i,\overline{b}_i)} \\
&\quad \text{aufgrund der Dreiecksungleichung} \\
&\quad \text{d.h., } w(\overline{a}_i,\overline{b}_i) \leq w(\overline{a}_i,\overline{c}_i)+w(\overline{c}_i,\overline{b}_i) \\
&\leq \underbrace{w(\overline{a},\overline{c})}_{\leq w(s')} + \underbrace{w(\overline{c}_i,\overline{b}_i)}_{=w(s_{n+1})} \\
&\leq w(s')+w(s_{n+1}) = w(s)
\end{aligned}
$$

Beziehung zwischen Edit- und Alignment-Distanz

Beweis.

Den Fall $\overline{a}_i = \overline{b}_i = -$ muss man gesondert betrachten.
Hier wird das verbotene Alignment von Leerzeichen im Alignment $(\overline{a}, \overline{b})$ eliminiert:

$$\begin{aligned} w(\overline{a}, \overline{b}) &= w(\overline{a}, \overline{c}) - w(\overline{a}_i, \overline{c}_i) \\ &\leq w(\overline{a}, \overline{c}) + w(\overline{c}_i, \overline{b}_i) \\ &\leq w(s') + w(s_{n+1}) = w(s) \end{aligned}$$

Beziehung zwischen Edit- und Alignment-Distanz

Beweis.

- Es bleibt noch der Fall, wenn $s_{n+1} = (-, y)$ mit $y \in \Sigma$ eine Insertion ist.
- Dann erweitern wir das Alignment $(\overline{a}, \overline{c})$ von a und c zu einem eigentlich *unzulässigen Alignment* $(\overline{a'}, \overline{c'})$ von a und c wie folgt.
- Es gibt ein $i \in [0 : |b|]$ mit $b_i = y$ und $b = c_1 \cdots c_i \cdot y \cdot c_{i+1} \cdots c_{|a|}$.
- Sei j die Position, nach der das Symbol y in $\overline{c}$ eingefügt wird.
- Dann setzen wir $\overline{a} = \overline{a}_1 \cdots \overline{a}_j \cdot - \cdot \overline{a}_{j+1} \cdots \overline{a}_{|\overline{c}|}$, $\overline{c} = \overline{c}_1 \cdots \overline{c}_j \cdot - \cdot \overline{c}_{j+1} \cdots \overline{c}_{|\overline{c}|}$ und $\overline{b} = \overline{c}_1 \cdots \overline{c}_j \cdot y \cdot \overline{c}_{j+1} \cdots \overline{c}_{|\overline{c}|}$.

Beziehung zwischen Edit- und Alignment-Distanz

Beweis.

$\overline{a}$		$-$	
$\overline{c}$		$-$	
$\overline{b}$		y	

$\overline{a}|_\Sigma = a$
$\overline{c}|_\Sigma = c$
$\overline{b}|_\Sigma = b$

Skizze: $s_{n+1} = (x, y)$ ist eine Insertion

- $(\overline{a}, \overline{c})$ ist jetzt wegen der Spalte $(-, -)$ kein Alignment mehr.
- jedoch nur noch interessant: Alignment $(\overline{a}, \overline{b})$ von a und b

$$\begin{aligned} w(\overline{a}, \overline{b}) &= w(\overline{a}, \overline{c}) + w(-, y) \\ &\quad \text{Nach Induktionsvoraussetzung} \\ &\leq w(s') + w(s_{n+1}) = w(s) \end{aligned}$$

Beziehung zwischen Edit- und Alignment-Distanz

Also ist die Alignment-Distanz durch die Edit-Distanz beschränkt, falls die zugrunde liegende Kostenfunktion eine Metrik ist:

Folgerung

Ist $w : \overline{\Sigma}_0^2 \to \mathbb{R}_+$ *eine metrische Kostenfunktion, dann gilt für alle* $a, b \in \Sigma^*$*:*

$$\overline{d}_w(a, b) \leq d_w(a, b).$$

Zusammengefasst ergibt sich für den Fall einer metrischen Kostenfunktion die Gleichheit von Edit- und Alignment-Distanz:

Theorem

Ist w *eine Metrik, dann gilt für* $a, b \in \Sigma^*$*:* $d_w(a, b) = \overline{d}_w(a, b)$.

Definitheit

- Alle drei Bedingungen der Metrik werden wirklich benötigt.
- Insbesondere für die Definitheit mache man sich klar, dass man zumindest auf die Gültigkeit von $w(x, x) = 0$ für alle $x \in \Sigma$ nicht verzichten kann.

- Manchmal will man aber auf die Definitheit verzichten, d.h. manche Zeichen sollen gleicher als andere sein.
- Dann schwächt man die Bedingungen an die Kostenfunktion $w : \overline{\Sigma}_0^2 \to \mathbb{R}_+$ wie folgt ab.

Definitheit

Definition

Eine Kostenfunktion $w : \overline{\Sigma}_0^2 \to \mathbb{R}_+$ für ein Distanzmaß heißt sinnvoll wenn folgende Bedingungen erfüllt sind:

(D1) $\forall x, y \in \Sigma : \; w(x, x) \leq w(x, y)$;

(D2) $\forall x, y \in \Sigma : w(x, x) \leq 2w(y, -)$;

(D3) $\forall x, y \in \overline{\Sigma} : w(x, y) = w(y, x)$;

(D4) $\forall x, y, z \in \overline{\Sigma} : w(x, z) \leq w(x, y) + w(y, z)$.

In der obigen Definition sind die (Un)gleichungen für $x, y, z \in \overline{\Sigma}$ nur dann gültig, wenn maximal ein Leerzeichen involviert ist.

Definitheit

- Von der Definitheit bleibt hier also nur noch übrig, dass gleiche Zeichen einen geringeren Abstand (aber nicht notwendigerweise 0) haben müssen als verschiedene Zeichen (D1).
- Damit in einem Alignment ein Match (x, x) nicht durch eine Deletion $(x, -)$ und eine Insertion von $(-x)$ von x ersetzen kann, wurde D2 eingeführt.
- D2 ist sogar noch etwas stärker, da es so in den beiden folgenden Beweisen benötigt wird.
- Will man die Äquivalenz von Edit- und Alignment-Distanz weiterhin sicherstellen, so wird man ebenfalls $\max\{w(x, x) : x \in \Sigma\} = 0$ fordern.
- Im Folgenden werden wir für alle Kostenfunktionen für Distanzmaße voraussetzen, dass zumindest D1 bis D4 gelten.

Ähnlichkeitsmaße

- Manchmal ist der Begriff der Ähnlichkeit von zwei Zeichen angemessener als der Begriff der Unterschiedlichkeit.
- Im Unterschied zu Distanzen werden nun gleiche Zeichen mit einem positiven Gewicht 'belohnt', während ungleiche Zeichen mit einen negativen Gewicht oder einem vergleichsweise kleinen positiven Gewicht 'bestraft' werden.
- Insbesondere besteht die Möglichkeit, die Gleichheit von gewissen Zeichen stärker zu bewerten als die von anderen Zeichenpaaren.

Kostenfunktion

Definition

Eine Kostenfunktion $w' : \overline{\Sigma}_0^2 \to \mathbb{R}$ für ein Ähnlichkeitsmaß heißt sinnvoll, wenn die folgenden Bedingungen erfüllt sind:

(S1) $\forall x \in \Sigma : w'(x,x) \geq 0 \ \wedge \ w'(x,-) \leq 0$

(S2) $\forall x,y \in \Sigma : w'(x,x) \geq w'(x,y)$

(S3) $\forall x,y \in \overline{\Sigma} : w'(x,y) = w'(y,x)$

(S4) $\forall x,y \in \Sigma : w'(x,y) \geq w'(x,-) + w'(-,y)$

In der obigen Definition sind die (Un)gleichungen für $x, y \in \overline{\Sigma}$ nur dann gültig, wenn maximal ein Leerzeichen involviert ist.

Kostenfunktion

- Positive Werte beschreiben Ähnlichkeiten und negative Werte Unähnlichkeiten. Ein Zeichen sollte zu sich selbst zumindest keinen negativen Ähnlichkeitswert haben. Dagegen sollten InDels keinen positiven Ähnlichkeitswert haben.
- In jedem Fall sollte ein Zeichen zu sich selbst mindestens so ähnlich sein wie zu einem anderen Zeichen (S2).
- Die Ähnlichkeit soll symmetrisch sein (S3).
- Würde S4 nicht gelten, so wäre es in einem Alignment besser, die Substitution (x, y) durch eine Deletion von x und eine Insertion von y zu ersetzen. Dies macht meist keinen Sinn, kann aber in einigen Fällen sinnvoll sein. (Für die folgenden Lemmata ist diese Eigenschaft jedoch beweistechnisch notwendig.)

Alignment-Ähnlichkeit

Im Folgenden werden wir für alle Kostenfunktionen für Ähnlichkeitsmaße mindestens voraussetzen, dass S1 und S2 gelten.

Definition

Sei $w' : \overline{\Sigma}_0^2 \to \mathbb{R}$ eine Kostenfunktion und sei $(\overline{a}, \overline{b})$ ein Alignment für $a, b \in \Sigma^*$. Dann ist die Ähnlichkeit des Alignments von $(\overline{a}, \overline{b})$ definiert als:

$$w'(\overline{a}, \overline{b}) := \sum_{i=1}^{|\overline{a}|} w'(\overline{a_i}, \overline{b_i})$$

Die Alignment-Ähnlichkeit von $a, b \in \Sigma^*$ ist definiert als:

$$s(a, b) := \max \left\{ w'(\overline{a}, \overline{b}) \ : \ (\overline{a}, \overline{b}) \text{ ist Alignment für } a, b \right\}$$

Beziehung zwischen Distanz- und Ähnlichkeitsmaßen

Lemma

Sei $w : \overline{\Sigma}_0^2 \to \mathbb{R}_+$ *eine sinnvolle Kostenfunktion für ein Distanzmaß. Dann existiert ein* $C \in \mathbb{R}_+$*, so dass* w' *mit*

$$\begin{aligned} w'(a,b) &= C - w(a,b) \\ w'(a,-) &= \frac{C}{2} - w(a,-) \end{aligned}$$

für alle $a, b \in \Sigma$ *eine sinnvolle Kostenfunktion für ein Ähnlichkeitsmaß ist.*

Beziehung zwischen Distanz- und Ähnlichkeitsmaßen

Beweis.

Sei $C := \max\{w(x,x) : x \in \Sigma\} \geq 0$. Dann gilt für alle $a \in \Sigma$

$$\begin{aligned} w'(a,a) &= C - w(a,a) \\ &= \max\{w(x,x) : x \in \Sigma\} - w(a,a) \geq 0 \end{aligned}$$

sowie mit D2

$$\begin{aligned} w'(a,-) &= \frac{C}{2} - w(a,-) \\ &= \frac{1}{2}\left(\max\{w(x,x) : x \in \Sigma\} - 2w(a,-)\right) \overset{D2}{\leq} 0 \end{aligned}$$

und somit S1.

Beziehung zwischen Distanz- und Ähnlichkeitsmaßen

Beweis.

Für jedes $a \in \Sigma$ und jedes $a \neq b \in \Sigma$ gilt aufgrund von D1 $w(a,a) \leq w(a,b)$ und somit:

$$w'(a,a) = C - w(a,a) \geq C - w(a,b) = w'(a,b).$$

Damit haben wir die Eigenschaft S2 nachgewiesen.
Die Eigenschaft S3 folgt aus der Definition von w' und D3.
Für den Nachweis von S4 gilt für $a, b \in \Sigma$ mit Hilfe der Dreiecksungleichung:

$$\begin{aligned} w'(a,b) &= C - w(a,b) \\ &\geq C - w(a,-) - w(-,b) = w'(a,-) + w'(b,-). \end{aligned}$$

Damit haben wir die Eigenschaft S4 nachgewiesen. □

Beziehung zwischen Distanz- und Ähnlichkeitsmaßen

Lemma

Sei $w' : \overline{\Sigma}_0^2 \to \mathbb{R}_+$ *eine sinnvolle Kostenfunktion für ein Ähnlichkeitsmaß. Dann existiert ein* $D \in \mathbb{R}_+$*, so dass* w *mit*

$$\begin{aligned} w(a,b) &= D - w'(a,b) \\ w(a,-) &= \frac{D}{2} - w'(a,-) \end{aligned}$$

für alle $a, b \in \Sigma$ *eine sinnvolle Kostenfunktion für ein Distanzmaß ist.*

Beziehung zwischen Distanz- und Ähnlichkeitsmaßen

Beweis.

Wir wählen
$A = \max\{w'(x,y) + w'(y,z) - w'(x,z) : x,y,z \in \overline{\Sigma}\} \geq 0$ und
$B = \max\{w'(x,y) : x,y \in \Sigma\} \geq 0$.
$D = \max\{A,B\} \geq 0$.

Nach Wahl von D gilt dann für $a,b \in \Sigma$:

$$\begin{aligned} w(a,b) &= D - w'(a,b) \\ &\geq \max\{w'(x,y) : x,y \in \Sigma\} - w'(a,b) \geq 0. \end{aligned}$$

Wegen S1 gilt $w'(a,-) \leq 0$ und daher ist

$$w(a,-) = \frac{D}{2} - w'(a,-) \geq 0$$

Somit haben wir nachgewiesen, dass $w : \overline{\Sigma} \times \overline{\Sigma} \to \mathbb{R}_+$.

Beziehung zwischen Distanz- und Ähnlichkeitsmaßen

Beweis.

D1: Wegen S2 gilt für alle $a, b \in \Sigma$: $w'(a,a) \geq w'(a,b)$ und somit:

$$\begin{aligned} w(a,a) &= D - w'(a,a) \\ &\leq D - w'(a,b) = w(a,b). \end{aligned}$$

D2: Nach S1 gilt für alle $a, b \in \Sigma$:

$$w'(a,a) \geq 0 \geq w'(b,-) \geq 2w'(b,-)$$

Also:

$$\begin{aligned} w(a,a) &= D - w'(a,a) \\ &\leq D - 2w'(b,-) = 2w(b,-). \end{aligned}$$

D3: folgt nach Definition von w unmittelbar aus S3.

Beziehung zwischen Distanz- und Ähnlichkeitsmaßen

Beweis.

D4: (d.h. $w(a,c) \leq w(a,b) + w(b,c)$ für alle $a, b, c \in \overline{\Sigma}$)

1. Fall: $a, b, c \in \Sigma$:

$$\begin{aligned} w(a,c) &= D - w'(a,c) \\ w(a,b) + w(b,c) &= 2D - w'(a,b) - w'(b,c) \end{aligned}$$

Also zu zeigen: $w'(a,b) + w'(b,c) - w'(a,c) \leq D$

Nach Definition von A gilt jedoch:

$$w'(a,b) + w'(b,c) - w'(a,c) \leq A \leq D$$

Beziehung zwischen Distanz- und Ähnlichkeitsmaßen

Beweis.

Es bleiben die Fälle, in denen genau ein Leerzeichen involviert ist.

2. Fall: $c = -$ und $a, b \in \Sigma$: (analog: $a = -$ und $b, c \in \Sigma$)

$$\begin{aligned} w(a,-) &= D/2 - w'(a,-) \\ w(a,b) + w(b,-) &= 3D/2 - w'(a,b) - w'(b,-) \end{aligned}$$

Also zu zeigen: $w'(a,b) + w'(b,-) - w'(a,-) \leq D$

Nach Definition von A gilt jedoch

$$w'(a,b) + w'(b,-) - w'(a,-) \leq A \leq D$$

Beziehung zwischen Distanz- und Ähnlichkeitsmaßen

Beweis.

3. Fall: $b = -$ und $a, c \in \Sigma$:

$$\begin{aligned} w(a,c) &= D - w'(a,c) \\ w(a,-) + w(c,-) &= D - w'(a,-) - w'(c,-) \end{aligned}$$

Also zu zeigen: $w'(a,c) \geq w'(a,-) + w'(c,-)$

Dies ist aber gerade die Bedingung S4. □

Damit haben wir trotz der unterschiedlichen Definition gesehen, dass Distanzen und Ähnlichkeiten eng miteinander verwandt sind.