Universität Bielefeld

Genominformatik
Technische Fakultät

Anzahl globaler Alignments

APPENDIX B

Pairwise Sequence Alignment (Extended Material)

B.1 The Number of Global Alignments

Roland Wittler

13. April 2015

Motivation

Können wir alle globalen Alignments zweier Sequenzen aufzählen?

```
AB    -AB    AB-
CD    CD-    -CD
```

```
AB-   A-B      -AB   A-B
C-D   CD-      C-D   -CD
```

```
--AB   -A-B   A--B   A-B-   -AB-   AB--
CD--   C-D-   -CD-   -C-D   C--D   --CD
```

$N(2,2) = 13$ $\boxed{N'(2,2) = 6}$

Übersicht

Übersicht

1. Motivation
2. **Algorithmischer Ansatz**
3. Kombinatorischer Ansatz
4. Vernachlässigung der Reihenfolge von InDels
5. Zusammenfassung

Algorithmischer Ansatz

Anzahl an Möglichkeiten einen leeren String mit einem anderen zu alignieren: $N(m,0) = N(0,n) = 1$

Match/Mismatch: $N(m–1,n–1)$ Möglichkeiten die Präfixe zu alignieren

$$\left(\begin{array}{l} \overbrace{.\ \ .\ \ -\ \ .\ \ .\ \ .\ \ .\ \ .}^{m\text{-}1}\ \ \mathtt{a} \\ \underbrace{.\ \ .\ \ .\ \ .\ \ -\ \ -\ \ .\ \ .}_{n\text{-}1}\ \ \mathtt{b} \end{array}\right)$$

Algorithmischer Ansatz

Anzahl an Möglichkeiten einen leeren String
mit einem anderen zu alignieren: $N(m,0) = N(0,n) = 1$

Match/Mismatch: $N(m–1,n–1)$ Möglichkeiten
die Präfixe zu alignieren

Deletion: $N(m–1,n)$ Möglichkeiten ...

$$\left(\begin{array}{cc} \overbrace{\texttt{. . - . . . . .}}^{m-1} & \texttt{a} \\ \underbrace{\texttt{. . . . - - . .}}_{n} & \texttt{-} \end{array}\right)$$

Algorithmischer Ansatz

Anzahl an Möglichkeiten einen leeren String
mit einem anderen zu alignieren: $N(m,0) = N(0,n) = \ 1$

Match/Mismatch: $N(m-1,n-1)$ Möglichkeiten
die Präfixe zu alignieren

Deletion: $N(m-1,n)$ Möglichkeiten ...

Insertion: $N(m,n-1)$ Möglichkeiten ...

Rekursion: $N(m,n) = N(m-1,n-1) + N(m-1,n) + N(m,n-1)$

Algorithmischer Ansatz

1	1	1	1	1	...
1	3	5	7	9	...
1	5	13	25	41	...
1	7	25	63	...	
1	9	41	...		
....	...	...			

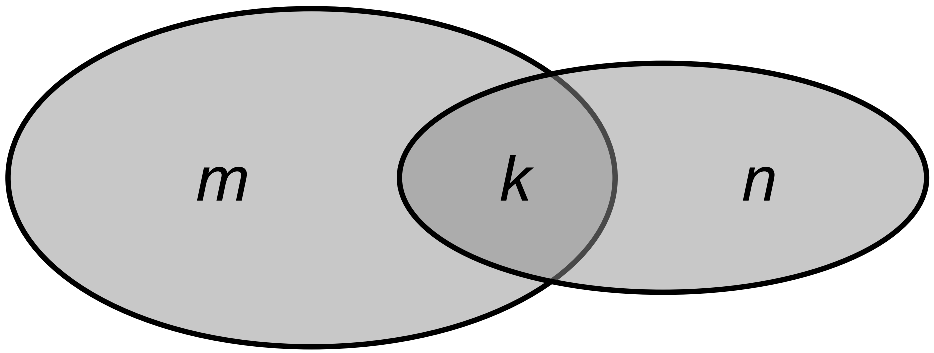
Übersicht

1. Motivation
2. Algorithmischer Ansatz
3. **Kombinatorischer Ansatz**
4. Vernachlässigung der Reihenfolge von InDels
5. Zusammenfassung

Kombinatorischer Ansatz

$N(m,n)$ symmetrisch ➔ Annahme: $n \leq m$

Anzahl k an Match/Mismatch-Spalten: $0 \leq k \leq n$



➔ $n - k$ Insertionen, $m - k$ Deletionen

➔ $n + m - k$ *Spalten insgesamt*

Kombinatorischer Ansatz

$\binom{m+n-k}{k}$ Möglichkeiten Match/Mismatch-Spalten zu positionieren

$\binom{m+n-2k}{n-k}$ Möglichkeiten InDel-Spalten zu positionieren

$r := n - k$

$n - k$ Insertionen, $m - k$ Deletionen

$n + m - k$ Spalten insgesamt

$$N(m,n) = \sum_{r=0}^{n} \binom{m+r}{n-r} \cdot \binom{m-n+2r}{r}$$

Kombinatorischer Ansatz

Werte auf der Diagonalen:

$$N(n,n) = \sum_{r=0}^{n} \binom{n+r}{n-r} \cdot \binom{2r}{r}$$

$$\approx (1 / \sqrt{n})\,(1 + \sqrt{2})^{2n+1}$$ ➔ exponentiell

$$N(m,n) = \sum_{r=0}^{n} \binom{m+r}{n-r} \cdot \binom{m-n+2r}{r}$$

Übersicht

1. Motivation
2. Algorithmischer Ansatz
3. Kombinatorischer Ansatz
4. **Vernachlässigung der Reihenfolge von InDels**
5. Zusammenfassung

Vernachlässigung der Reihenfolge von InDels

```
AB      -AB      AB-
CD      CD-      -CD
```

```
AB-    A-B      -AB    A-B
C-D    CD-      C-D    -CD
```

```
--AB    -A-B    A--B    A-B-    -AB-    AB--
CD--    C-D-    -CD-    -C-D    C--D    --CD
```

$N'(2,2) = 6$

Vernachlässigung der Reihenfolge von InDels

$\binom{n}{k}$ Möglichkeiten Match/Mismatch-Positionen in kurzer Seq.

$\binom{m}{k}$ Möglichkeiten Match/Mismatch-Positionen in langer Seq.

$$N'(m,n) = \sum_{k=0}^{n} \binom{n}{k} \cdot \binom{m}{k} = \binom{m+n}{k}$$

$$N'(n,n) = \binom{2n}{k} \approx \frac{4^n}{\sqrt{(\pi n)}} \, exp(-3/24n)$$ ➔ exponentiell

Übersicht

1. Motivation
2. Algorithmischer Ansatz
3. Kombinatorischer Ansatz
4. Vernachlässigung der Reihenfolge von InDels
5. **Zusammenfassung**

Zusammenfassung

Rekursion: $N(m,n) = N(m-1,n-1) + N(m-1,n) + N(m,n-1)$

$$N(m,n) = \sum_{r=0}^{n} \binom{m+r}{n-r} \cdot \binom{m-n+2r}{r}$$ ➔ exponentiell

$$N(n,n) \approx (1 / \sqrt{n}) (1 + \sqrt{2})^{2n+1}$$ ➔ exponentiell

$$N'(m,n) = \binom{m+n}{k}$$ ➔ exponentiell

$$N'(n,n) \approx \frac{4^n}{\sqrt{(\pi n)}} \; exp(-3/24n)$$ ➔ exponentiell