Platzkomplexität eines Bioinformatik-Problems

Haplotypisierung mittels perfekter Phylogenien

Sommerakademie Rot an der Rot — AG 1
Wieviel Platz brauchen Algorithmen wirklich?

Sebastian Dörner

Fakultät für Informatik
Otto-von-Guericke-Universität Magdeburg

18. August 2010

Überblick

1. Begriffe aus Biologie und Genetik
2. Das Problem der Haplotypisierung
   - Sichtweise eines Biologen
   - Sichtweise eines Informatikers
3. Lösungsansatz: Perfekte Phylogenien
4. Haplotypisierung in LOGSPACE
   - Charakterisierung perfekter Phylogenien
   - Auflösung von 2-Einträgen im Genotyp
   - Auflösungsgraphen
   - Reduktion von PPH auf BIPARTITION

Begriffe aus Biologie und Genetik

- Basiseinheiten der DNA: Adenin **A**, Guanin **G**, Thymin **T**, Cytosin **C**
- Chromosom: . . . CTTTGAAGGGAATTAAA. . .
- Genotyp: Erbbild eines Organismus, bestehend aus Chromosomen

  Haplotyp . . . CTTTGAAGGGAATTAAA. . .
  . . . CTTTGAAGGGGATTAAA. . . } Genotyp (diploid)
- **S**ingle **N**ucleotide **P**olymorphism

Sichtweise eines Biologen

- große Teile der DNA zwischen Individuen gleich
- $\Rightarrow$ Untersuchung von einzelnen Basenpaaren, die häufig variieren (SNPs)
- chem. Untersuchung an DNA (diploid): an bestimmtem Ort liegen bestimmte Basen vor

  ...CTTTGAAGGGAATTAAA...
  ...CTTTGAAGGGGATTAAA...
  ↑
  - gleiche Basen auf beiden Chromosomen: vollständiger Aufbau bekannt
  - unterschiedliche Basen: Welche Base liegt auf welchem Chromosom?

Sichtweise eines Informatikers

- Haplotypen $h \in \{0,1\}^m$
- Genotyp g als Konkatenation der Mengen $\{h[i], h'[i]\}, 1 \leq i \leq m$
  einfachere Kodierung: $\{0\} \to 0, \{1\} \to 1, \{0,1\} \to 2 \Rightarrow g \in \{0,1,2\}^n$

Sichtweise eines Informatikers

- Haplotypen $h \in \{0,1\}^m$
- Genotyp g als Konkatenation der Mengen $\{h[i], h'[i]\}, 1 \leq i \leq m$
  einfachere Kodierung: $\{0\} \to 0, \{1\} \to 1, \{0,1\} \to 2 \Rightarrow g \in \{0,1,2\}^n$
- Beispiel:
  $h = 0100, h' = 0111 \Rightarrow g = 0122$
  $h = 0101, h' = 0110 \Rightarrow g = 0122$
  - $h[i] = h[j] \neq h'[i] = h'[j]$: h und h' werden in i und j gleich aufgelöst
  - $h[i] = h'[j] \neq h'[i] = h[j]$: h und h' werden in i und j ungleich aufgelöst
- Haplotypmatrix/Genotypmatrix:

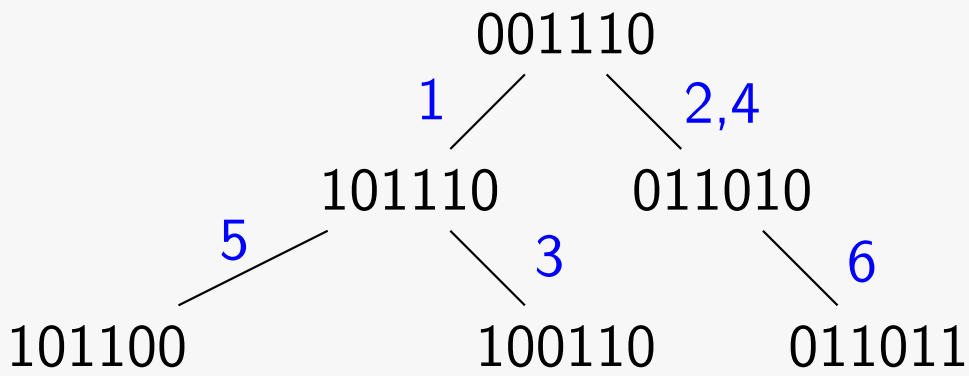
$$\begin{pmatrix} 010010 \\ 011011 \\ 110100 \\ 010101 \end{pmatrix} \Longrightarrow \begin{pmatrix} 012012 \\ 210102 \end{pmatrix}$$

Lösungsansatz: Perfekte Phylogenien

- Annahmen:
  - Haplotypen haben gemeinsame "Vorfahren"
  - zwischen Generationen ändern sich die Basen an den SNP-Orten
  - am gleichen Ort ändert sich die Base nur ein mal

$\Rightarrow$ Anordnung in einer perfekten Phylogenie

$$\begin{pmatrix} 100110 \\ 001110 \\ 011011 \\ 101110 \\ 011010 \\ 101100 \end{pmatrix}$$



Lösungsansatz: Perfekte Phylogenien

- Haplotypmatrix B lässt eine perfekte Phylogenie zu, wenn Wurzelbaum T existiert mit
  1. Jede Zeile von B beschriftet genau einen Knoten von T
  2. Jede Spalte von B beschriftet genau eine Kante von T und jede Kante wird von mindestens einer Spalte beschriftet
  3. Für alle Paare (h, h') von Zeilen aus B und jede Spalte i gilt $h[i] \neq h'[i]$ genau dann, wenn i auf dem Pfad von h zu h' liegt
- B lässt eine gerichtete perfekte Phylogenie zu, wenn B erweitert um den 0-Haplotypen eine perfekte Phylogenie zulässt

1 Begriffe aus Biologie und Genetik

2 Das Problem der Haplotypisierung
- Sichtweise eines Biologen
- Sichtweise eines Informatikers

3 Lösungsansatz: Perfekte Phylogenien

4 Haplotypisierung in LOGSPACE
- Charakterisierung perfekter Phylogenien
- Auflösung von 2-Einträgen im Genotyp
- Auflösungsgraphen
- Reduktion von PPH auf BIPARTITION

Charakterisierung perfekter Phylogenien

- Induzierte Menge $\text{ind}^B(i,j)$ zweier Spalten i und j der Haplotypmatrix B enthält alle Zeichenketten aus $\{00, 01, 10, 11\}$, die in den Spalten i und j vorkommen

$$B = \begin{pmatrix} 010010 \\ 011011 \\ 110100 \\ 010101 \end{pmatrix} \Longrightarrow \text{ind}^B(2,4) = \{10, 11\}$$

Charakterisierung perfekter Phylogenien

- Induzierte Menge $\text{ind}^B(i,j)$ zweier Spalten i und j der Haplotypmatrix B enthält alle Zeichenketten aus $\{00,01,10,11\}$, die in den Spalten i und j vorkommen

$$B = \begin{pmatrix} 010010 \\ 011011 \\ 110100 \\ 010101 \end{pmatrix} \Longrightarrow \text{ind}^B(2,4) = \{10,11\}$$

- four gamete property: B lässt eine perfekte Phylogenie zu gdw. $\forall i \forall j : \{00,01,10,11\} \neq \text{ind}^B(i,j)$
- three gamete property: B lässt eine ger. perf. Phylogenie zu gdw. $\forall i \forall j : \{01,10,11\} \notin \text{ind}^B(i,j)$

$\Rightarrow$ PPH = $\{A \mid A$ ist Genotyp-Matrix und erlaubt eine perfekte Phylogenie$\}$, analog DPPH

Auflösung von 2-Einträgen im Genotyp

- Induzierte Mengen der Genotypmatrix:
  $xy \in \text{ind}^A(i,j) \subseteq \{00,01,10,11\}$ gdw. A enthält Genotyp g mit
  $(g[i] = x \wedge g[j] = y) \vee (g[i] = x \wedge g[j] = 2) \vee (g[i] = 2 \wedge g[j] = y)$
- $\{01,10,11\} \subseteq \text{ind}^A(i,j) \subseteq \text{ind}^B(i,j) \Rightarrow A \notin$ DPPH
- betrachte Haplotypmatrizen, die die three gamete property erfüllen
- einfache Folgerungen für 2 Spalten
  - $\{01,10\} \subseteq \text{ind}^A(i,j) \Rightarrow$ alle Genotypen g mit $g[i] = g[j] = 2$ werden von den Haplotypen aus B in i und j ungleich aufgelöst
  - $\{11\} \subseteq \text{ind}^A(i,j) \Rightarrow$ Genotypen in i und j gleich aufgelöst

Auflösung unter Beteiligung mehrerer Spalten

- Genotyp g aus A mit $g[i] = g[j] = g[k] = 2$ und erklärenden Haplotypen h und h'
- $\{11\} \subseteq \text{ind}^A(i,j), \{01,10\} \subseteq \text{ind}^A(j,k)$

$\Rightarrow$ $h[i] = h[j] \neq h'[i] = h'[j]$, $h[j] = h'[k] \neq h'[j] = h[k]$

$\Rightarrow$ $h[i] = h'[k] \neq h'[i] = h[k] \Rightarrow$ ungleiche Auflösung in i und k

- komplexere Folgerungen für 3 Spalten können weitere Inferenzen nach sich ziehen

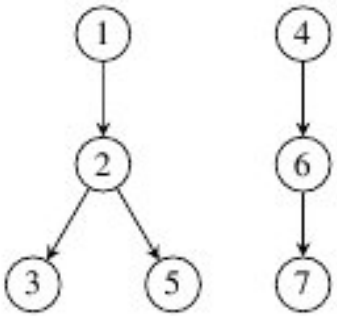
Auflösungsgraphen

- Graphen zur Modellierung gleicher und ungleicher Auflösung
  - Knoten im Graph $\widehat{=}$ Spalten der Genotypmatrix
  - Kanten mit Gewicht 0: gleiche Auflösung der Spalten inzidenter Knoten
  - Kanten mit Gewicht 1: ungleiche Auflösung der Spalten inzidenter Knoten
- konstruiere für jede Spalte i in A einen Auflösungsgraphen G_i
- G_i beschreibt Auflösungen für eine Menge A_i von Genotypen, wobei $A = \dot{\bigcup}_{1 \leq i \leq m} A_i$
- Auflösung eines Genotyps g: Betrachte nur G_i mit $g \in A_i$

Auflösungsgraphen: Konstruktion der A_i

- $i \succ^A j$ gdw. $\text{ind}^A(i,j) \subseteq \{00, 10, 11\}$ und die Spaltenvektoren i und j sind unterschiedlich
- $A_i = \{\text{Genotyp } g \text{ von A} \mid g[i] = 2 \ \wedge \ \forall j \neq i(g[j] = 2 \Rightarrow j \not\succ^A i) \ \wedge \ \forall j \neq i((g[j] = 2 \wedge \forall k \neq j(g[k] = 2 \Rightarrow k \not\succ^A j)) \Rightarrow j > i)\}$
- jeder Genotyp mit einem 2-Eintrag landet in genau einem A_i

1	2	3	4	5	6	7
0	0	0	1	0	1	0
1	2	2	0	0	0	0
2	2	2	2	0	0	0
1	2	0	0	2	0	0
0	0	0	2	0	2	0
0	0	0	2	0	2	2



$$A_1 = \{ \ 2\,2\,2\,2\,0\,0\,0 \ \}$$

$$A_2 = \left\{ \begin{array}{l} 1\,2\,2\,0\,0\,0\,0 \\ 1\,2\,0\,0\,2\,0\,0 \end{array} \right\}$$

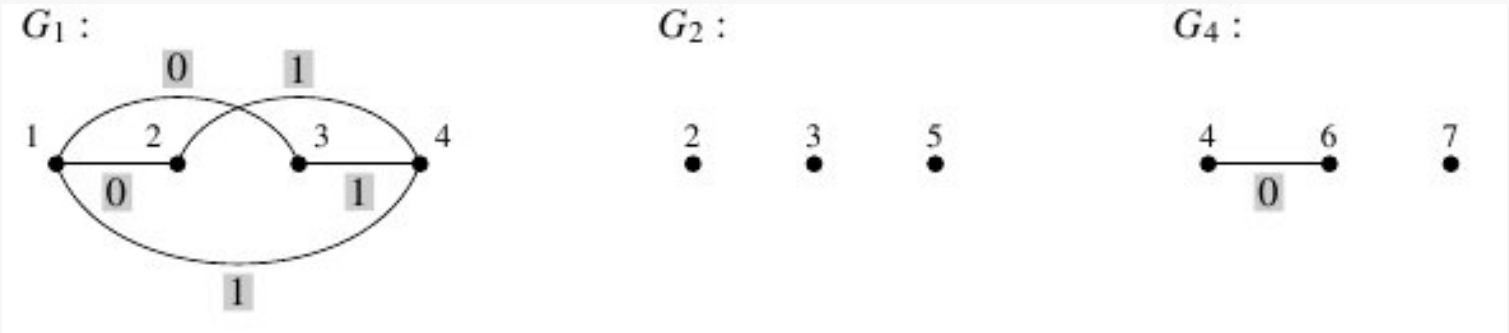
$$A_3 = \emptyset$$

$$A_4 = \left\{ \begin{array}{l} 0\,0\,0\,2\,0\,2\,0 \\ 0\,0\,0\,2\,0\,2\,2 \end{array} \right\}$$

$$A_5 = A_6 = A_7 = \emptyset$$

Auflösungsgraphen: Konstruktion der G_i

- konstruiere $G_i = (V_i, E_i)$ aus A_i
  - $k \in V_i \Leftrightarrow A_i$ enthält Genotyp mit einer 2 in Spalte k
  - Kantenmenge $E_i \subseteq \{\{k,l\} \mid k,l \in V_i\}$, Gewichte $w_i : E_i \to \{0,1\}$
  - $\{k,l\} \in E_i \wedge w_i(\{k,l\}) = 0$ gdw. $\exists g_1 \in A_i : g_1[k] = g_1[l] = 2$ und
    (a) $11 \in \text{ind}^A(k,l)$ oder
    (b) es gibt eine Spalte $j \neq i$ und $g_2 \in A_j$ mit $g_2[k] = g_2[l] = 2$
  - $\{k,l\} \in E_i \wedge w_i(\{k,l\}) = 1$ gdw. $\exists g_1 \in A_i : g_1[k] = g_1[l] = 2$ und $\{01,10\} \subseteq \text{ind}^A(k,l)$



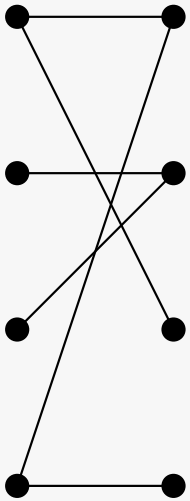
Auflösungsgraphen

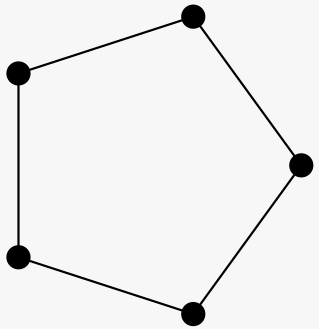
Lemma

Eine $n \times m$ Genotypmatrix A lässt eine gerichtete perfekte Phylogenie zu genau dann, wenn für jedes Paar $i, j \in \{1, \ldots, m\}$ gilt, dass $\{01, 10, 11\} \notin ind^A(i, j)$ und für alle $i \in \{1, \ldots, m\}$ der Graph G_i keinen Kreis mit ungeradem Gewicht enthält.

Reduktion von PPH auf BIPARTITION

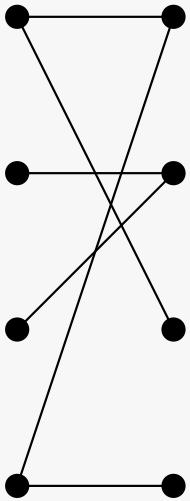
BIPARTITION: Menge aller bipartiten Graphen

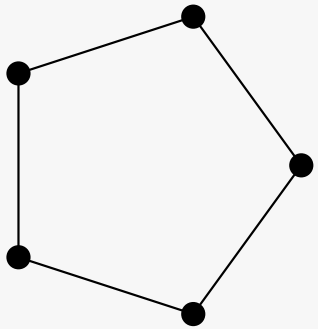


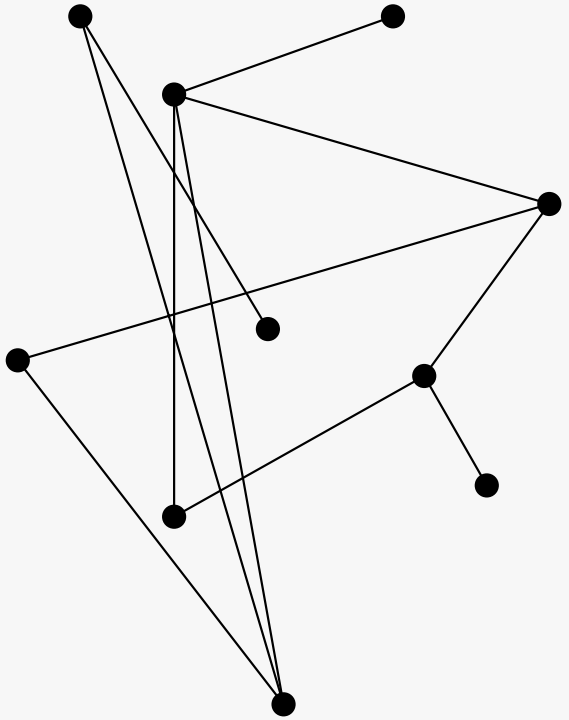


Reduktion von PPH auf BIPARTITION

BIPARTITION: Menge aller bipartiten Graphen







Reduktion von PPH auf BIPARTITION

- Kontruktives Vorgehen
  - Genotypmatrix $A \to A'$: Reduktion von PPH auf DPPH nach [EHK02]
  - konstruiere G: disjunkte Vereinigung aller Auflösungsgraphen von A'
  - ersetze in G Kanten mit Gewicht 0 durch 2 Kanten und lösche alle Kantengewichte

  $\to$ G', ermittle ob $G' \in$ BIPARTITION
- Korrektheit: $A \in$ PPH $\Leftrightarrow$ $A' \in$ DPPH $\Leftrightarrow$ G enhält keinen Kreis mit ungeradem Gewicht $\Leftrightarrow$ G' enthält keinen Kreis ungerader Länge $\Leftrightarrow$ G' ist bipartit

Mit BIPARTITION $\in L$ und PPH ist L-hart folgt

Theorem

PPH ist L-vollständig.

Literatur


Eleazar Eskin, Eran Halperin, and Richard M. Karp.
Efficient reconstruction of haplotype structure via perfect phylogeny.
Technical report, Berkeley, CA, USA, 2002.


Michael Elberfeld.
Perfect phylogeny haplotyping is complete for logspace.
CoRR, abs/0905.0602, 2009.


Jens Gramm, Arfst Nickelsen, and Till Tantau.
Fixed-parameter algorithms in phylogenetics.
Comput. J., 51(1):79–101, 2008.

Vielen Dank für die Aufmerksamkeit!