Platzkomplexität eines Bioinformatik-Problems

Haplotypisierung mittels perfekter Phylogenien

Sebastian Dörner

26. August 2010

Zusammenfassung

In der Bioinformatik besteht das Problem der Haplotypisierung darin, von bekannten Genotypen auf wahrscheinliche Haplotypen zu schließen. Eine von Gusfield beschriebene Methode dafür benötigt eine linear Speicherplatzgröße und basiert auf der Annahme, die vorkommenden Haplotypen seien in einer perfekten Phylogenie, einer Baumstruktur mit gewissen Eigenschaften, angeordnet. Der hier vorgestellte Algorithmus basiert auf der selben Annahme, benötigt aber nur logarithmischen Speicheraufwand. Da das Problem bekanntermaßen L-hart ist, folgt damit die L-Vollständigkeit der Haplotypisierung mittels perfekter Phylogenien.

Inhaltsverzeichnis

1 Begriffe aus Biologie und Genetik **1**

2 Das Problem der Haplotypisierung **2**
2.1 Sichtweise eines Biologen 2
2.2 Sichtweise eines Informatikers 2

3 Lösungsansatz: Perfekte Phylogenien **2**

4 Haplotypisierung in Logspace **3**
4.1 Charakterisierung perfekter Phylogenien 3
4.2 Auflösung von 2-Einträgen im Genotyp 3
4.3 Auflösungsgraphen 3
4.4 Reduktion von PPH auf BIPARTITION 5

1 Begriffe aus Biologie und Genetik

- Basiseinheiten der DNA (Nukleinsäuren): Adenin **A**, Guanin **G**, Thymin **T**, Cytosin **C**
- Chromosom: Strang von Nukleinsäuren
- Genotyp: Erbbild eines Organismus, bestehend aus Chromosomen
- diploider Genotyp: doppelter Chromosomensatz
- haploider Genotyp = Haplotyp: einzelner Chromosomensatz, auch als Teil eines diploiden Genotyps
- Allel: Ausprägung eines Gens an einem bestimmten Ort eines Chromosoms
- **S**ingle **N**ucleotide **P**olymorphism: Variation einzelner Basenpaare auf einem Chromosom

2 Das Problem der Haplotypisierung

2.1 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
  - gleiche Basen auf beiden Chromosomen: vollständiger Aufbau bekannt
  - unterschiedliche Basen: Welche Base liegt auf welchem Chromosom?

2.2 Sichtweise eines Informatikers

- Großteil der SNPs: nur zwei verschiedene Basenpaare $\Rightarrow$ Haplotypen $h \in \{0,1\}^m$
- Genotyp g der Haplotypen h und h' 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: $\begin{matrix} h = 0100, h' = 0111 \Rightarrow g = 0122 \\ h = 0101, h' = 0110 \Rightarrow g = 0122 \end{matrix}$ h und h' erklären jeweils g
  - $h[i] = h[j] \neq h'[i] = h'[j]$: 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): Jede Zeile ein Haplotyp (Genotyp)
- $2n \times m$ Haplotypmatrix B erklärt $n \times m$ Genotypmatrix A: für alle i erklären die Zeilen $2i-1$ und $2i$ von B den Genotyp in Zeile i von A
- Ziel: finde zu gegebener Genotypmatrix eine wahrscheinliche Haplotypmatrix

3 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 einem Wurzelbaum mit speziellen Eigenschaften (perfekte Phylogenie)

- 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

4 Haplotypisierung in Logspace

4.1 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

- 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

- PPH und DPPH sind nach [EHK02] aufeinander reduzierbar

4.2 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 $(g[i] = x \wedge g[j] = y) \vee (g[i] = x \wedge g[j] = 2) \vee (g[i] = 2 \wedge g[j] = y)$
- Für alle Haplotypmatrizen B, die A erklären: Wenn A keinen Genotyp mit 2 in Spalten i und j enthält: $\text{ind}^A(i, j) \subseteq \text{ind}^B(i, j)$

$\Rightarrow$ $\{01, 10, 11\} \subseteq \text{ind}^A(i, j) \Rightarrow A \notin$ 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
- komplexere Folgerungen für 3 Spalten können weitere Inferenzen nach sich ziehen

4.3 Auflösungsgraphen

- Modellierung gleicher und ungleicher Auflösung durch ungerichtete, kantengewichtete Graphen
  - 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 $A = \biguplus_{1 \leq i \leq m} A_i$
- Auflösung eines Genotyps g: Betrachte nur G_i mit $g \in A_i$
- $i >^A j$ gdw. $\text{ind}^A(i, j) \subseteq \{00, 10, 11\}$ und die Spaltenvektoren i und j sind unterschiedlich
- $A_i = \{$Genotyp g von A $\mid g[i] = 2 \ \wedge \ \forall j \neq i (g[j] = 2 \Rightarrow j \not>^A i) \ \wedge$
  $\forall j \neq i ((g[2] = 2 \wedge \forall k \neq j (g[k] = 2 \Rightarrow k \not>^A j)) \Rightarrow j > i)\}$
- jeder Genotyp mit einem 2-Eintrag landet in genau einem A_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
    $\{01, 10\} \subseteq \text{ind}^A(k, l)$

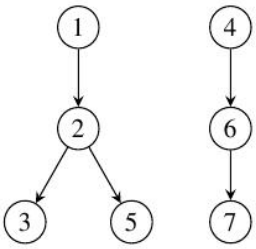
Lemma 4.1. *Eine $n \times m$ Genotypmatrix A lässt eine gerichtete perfekte Phylogenie zu genau dann, wenn für jedes Paar $i, j \in \{1, \ldots, m\}$ gilt, dass $\{01, 10, 11\} \notin ind^A(i, j)$ und für alle $i \in \{1, \ldots, m\}$ der Graph G_i keinen Kreis mit ungeradem Gewicht enthält.*

Beispiel:

Genotype matrix A:

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

Partial order $\succ^A$ on the columns of A:



Assignment of genotypes to sets A_i:

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

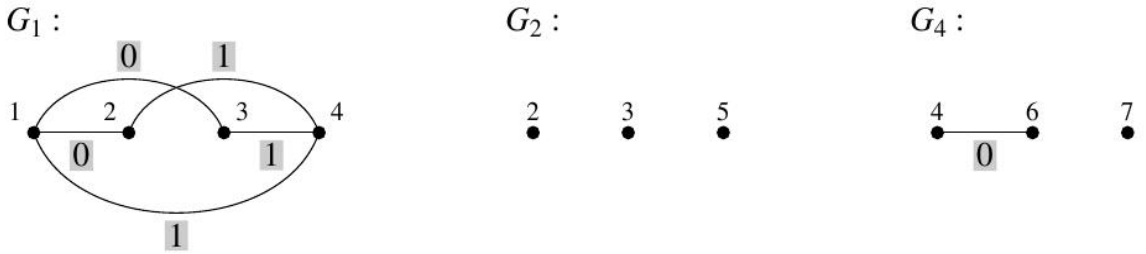
$$A_2 = \left\{ \begin{array}{c} 1\,2\,2\,0\,0\,0\,0 \\ 1\,2\,0\,0\,2\,0\,0 \end{array} \right\}$$

$$A_3 = \emptyset$$

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

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

Resolution graphs of columns 1, 2 and 4 with edge weights:



4.4 Reduktion von PPH auf BIPARTITION

- BIPARTITION: Menge aller bipartiten Graphen
- 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 1. *PPH ist L-vollständig.*

Literatur

[EHK02] Eleazar Eskin, Eran Halperin, and Richard M. Karp. Efficient reconstruction of haplotype structure via perfect phylogeny. Technical report, Berkeley, CA, USA, 2002.

[Elb09] Michael Elberfeld. Perfect phylogeny haplotyping is complete for logspace. *CoRR*, abs/0905.0602, 2009.

[GNT08] Jens Gramm, Arfst Nickelsen, and Till Tantau. Fixed-parameter algorithms in phylogenetics. *Comput. J.*, 51(1):79–101, 2008.