Markov Analysis of DNA Sequences

Elena Schmidt, Christina Wendel

Zielsetzung

- Grundlage: Arbeit über Markov Analyse des Lamda-Phagen Genoms
- Ziel: Bestimmung der Markov-Ordnung des Chromosoms 22 und Teilsequenzen des Chromosoms 22

Vorgehen

- Markov-Ordnung von Chromosom 22 bestimmen
- Ordnung einzelner Gene
- Einfluss der Sequenzlänge auf die Ordnung
- Ordnung nach Aufteilung des Chromosom in kleinere Abschnitte

Chromosom 22

- 34.748.584 Nucleotide
- 34,5 MB
- 844 Gene
- Chromosom von:

  ftp://ftp.sanger.ac.uk/pub/human/chr22/sequences/Chr_22/complete_sequence/
- Gene von:
  http://www.ncbi.nlm.nih.gov/genome/guide/human/

Markov-Ordnung k bestimmen

- Vorkommen k+1 langer Teilsequenzen zählen
- Wahrscheinlichkeiten für die Teilsequenzen schätzen:

  Bsp.: # AGC/(#AGA+ #AGC+ #AGG+ # AGT)

- für jede k+1 lange Sequenz log-Likelihood bestimmen
- BIC (Bayesian information criterion) berechnen:

  -2* logLikelihood +

  (# zu schätzender Parameter) * (# k+1-langer Sequenzabschnitte)
- k mit minimalem BIC (=maximaler Information) ist die gesuchte Markov-Ordnung (Annahme)

gesamtes Chromosom

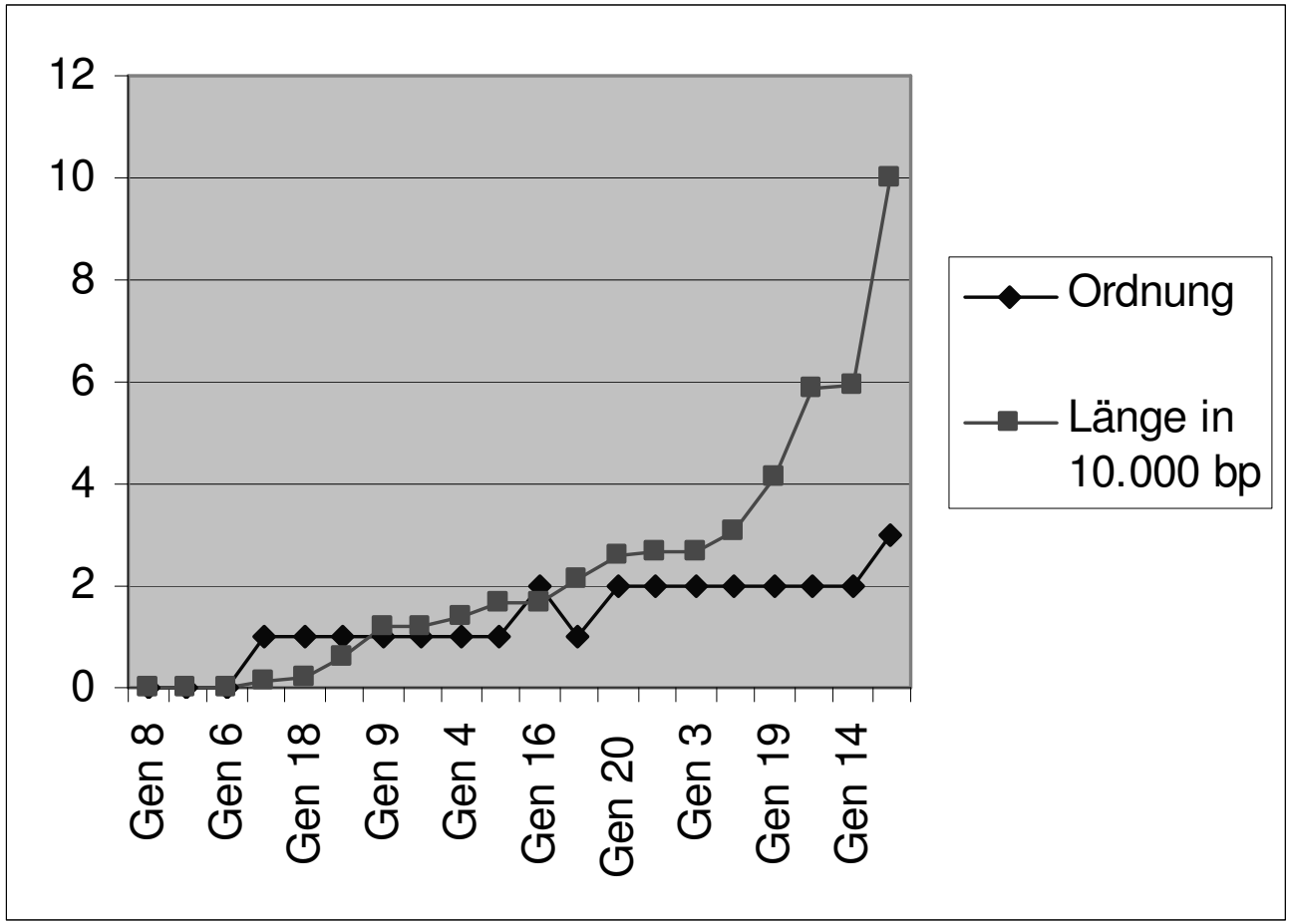
- Ergebnis: Markov-Ordnung 7
- mögliche Erklärungen:
  - „echte“ Abhängigkeiten
  - Verfälschung durch repetitive Sequenzen: ALUs: SINE- Familie, Länge 300bp

gesamtes Chromosom

- Teilsequenzen nach Häufigkeit sortiert
- Blast-Suche auf Alu-Datenbank

Teilsequenz	# in Chromosom 22	# in Alu-Sequenzen
CAGGCTGG	13856	129
CCAGCACT	6962	101
CAAGCGAT	2106	62
TTCGATCG	2	0

Auswertung von 20 Genen



Chromosom-Abschnitte

Markov-Ordnung	# 1.000er Abschnitte	# 20.000er Abschnitte
0	6798	34
1	27728	553
2	221	1137
3	2	9
4	0	5

Simulierte Sequenzen

- Verlässlichkeit des BIC
- Häufigkeitsverteilungen der gegebenen Sequenzen

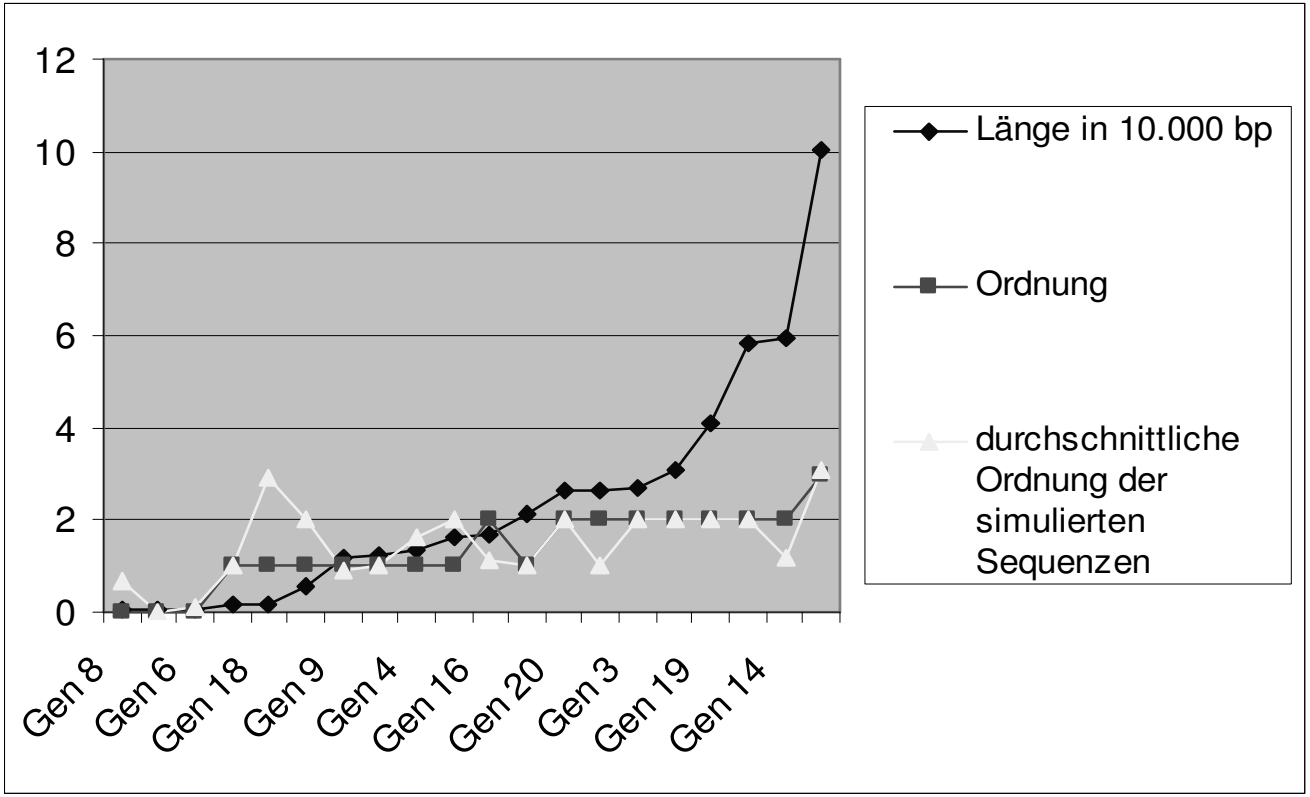
  ⟹ Erzeugung von Zufallssequenzen einer bestimmten Länge
- Bestimmung der Markov-Ordnung der Zufallssequenz

Simulierte Sequenzen (20 Gene)

- Für Gen der Länge n: Erzeugung einer Zufallssequenz der Länge n
- Ergebnis: 100% der Testsequenzen zeigten Übereinstimmung in der Markov-Ordnung

Simulierte Sequenzen (20 Gene)

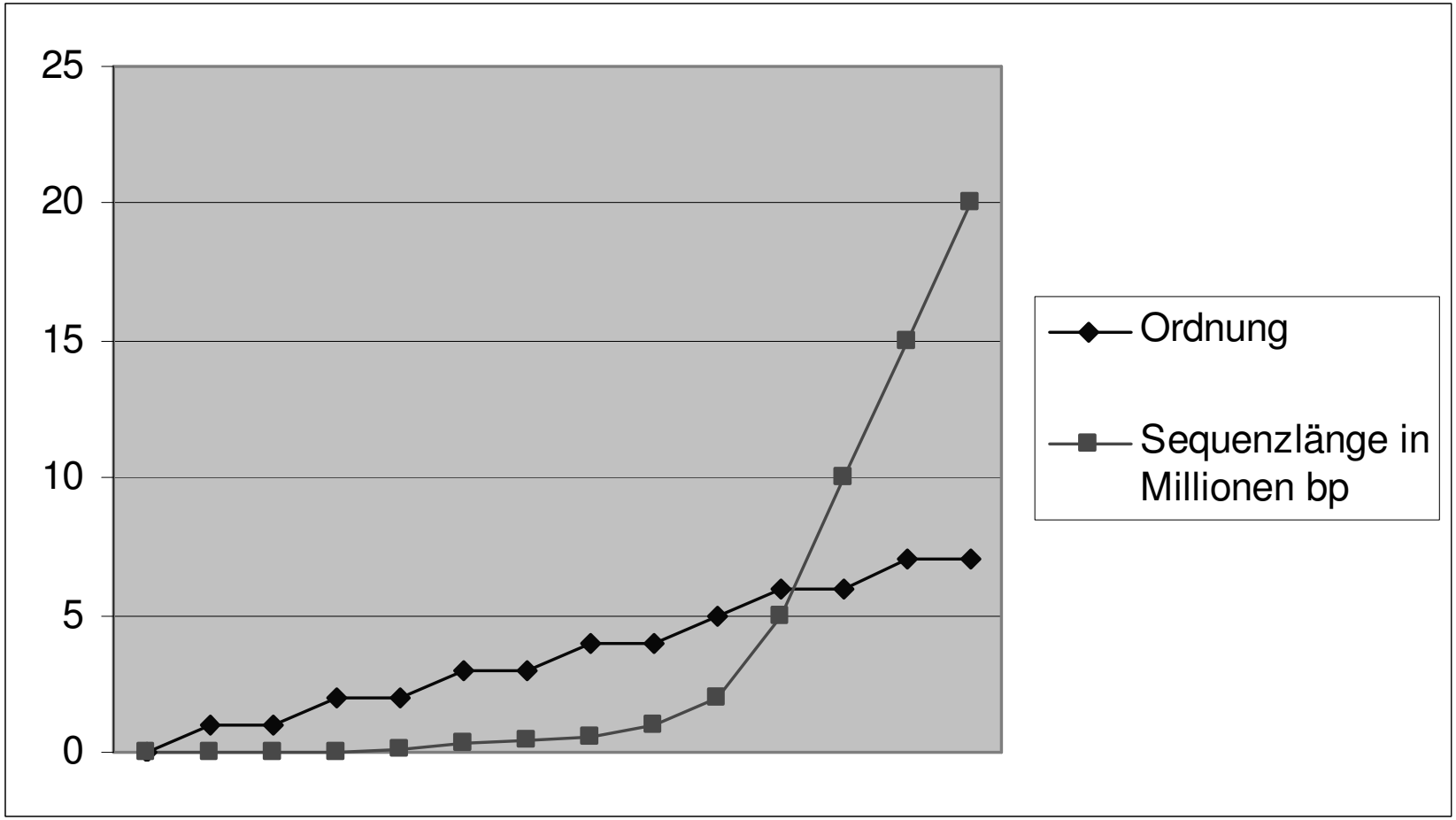
- mit relativen Häufigkeiten der nächsthöheren Ordnung



Simulierte Sequenzen (gesamtes Chromosom)

- Relative Häufigkeiten des Modells 7. Ordnung
- Erzeugen verschieden langer Sequenzen
- Markov-Ordnung bestimmen

Simulierte Sequenzen (gesamtes Chromosom)



Zusammenfassung

- nach unseren Beobachtungen hängt die Ordnung von der Sequenzlänge ab
- repetitive Sequenzen beeinflussen die Ordnung