Content Sensors

Michael Hanselmann

Hauptseminar Genvorhersage, SoSe 2005
Abteilung Theoretische Informatik

25. Oktober 2005

Einführung

- Zielsetzung: Gene in der DNA finden
- Schwierigkeiten:
  - codierende Bereiche von nicht codierenden unterscheiden
  - wo beginnt die Codierung eines Gens?

Einführung

- Zielsetzung: Gene in der DNA finden
- Schwierigkeiten:
  - codierende Bereiche von nicht codierenden unterscheiden
  - wo beginnt die Codierung eines Gens?
- verschiedene Verfahren, Ausgangsbasis immer eine DNA-Sequenz, Beispiel:

$$S = AGGACGGGATCA$$

mit

$$S_1 = A, S_2 = G, ..., S_l = G \text{ und Länge } l = 12$$

Einführung

- Zielsetzung: Gene in der DNA finden
- Schwierigkeiten:
  - codierende Bereiche von nicht codierenden unterscheiden
  - wo beginnt die Codierung eines Gens?
- verschiedene Verfahren, Ausgangsbasis immer eine DNA-Sequenz, Beispiel:

$$S = AGGACGGGATCA$$

mit

$$S_1 = A, S_2 = G, ..., S_l = G \text{ und Länge } l = 12$$

- eine solche Sequenz KANN Aminosäuren codieren

Codons

- 3er-Gruppe von Nukleotiden
- jedes Codon codiert eine bestimmte Aminosäure

Codons

- 3er-Gruppe von Nukleotiden
- jedes Codon codiert eine bestimmte Aminosäure
- es gibt 4^3=64 verschiedene Codons, die 20 Aminosäuren codieren

Aminosäure	Codon	Aminosäure	Codon
Arg	AGG	Met	ATG
Arg	AGA	Ile	ATA
Ser	AGT	Ile	ATT
Ser	AGC	Ile	ATC
Lys	AAG	Thr	ACG
Lys	AAA	Thr	ACA
Asn	AAT	Thr	ACT
Asn	AAC	Thr	ACC

- Einteilung von DNA-Sequenzen in Codons problematisch - drei mögliche Leserahmen:

$$S = \underbrace{AGG}_{C_1^1}\underbrace{ACG}_{C_2^1}\underbrace{GGA}_{C_3^1}\underbrace{TCA}_{C_4^1}$$

oder

$$S = A\underbrace{GGA}_{C_1^2}\underbrace{CGG}_{C_2^2}\underbrace{GAT}_{C_3^2}CA$$

oder

$$S = AG\underbrace{GAC}_{C_1^3}\underbrace{GGG}_{C_2^3}\underbrace{ATC}_{C_3^3}A$$

- Einteilung von DNA-Sequenzen in Codons problematisch - drei mögliche Leserahmen:

$$S = \underbrace{AGG}_{C_1^1}\underbrace{ACG}_{C_2^1}\underbrace{GGA}_{C_3^1}\underbrace{TCA}_{C_4^1}$$

oder

$$S = A\underbrace{GGA}_{C_1^2}\underbrace{CGG}_{C_2^2}\underbrace{GAT}_{C_3^2}CA$$

oder

$$S = AG\underbrace{GAC}_{C_1^3}\underbrace{GGG}_{C_2^3}\underbrace{ATC}_{C_3^3}A$$

- Codons kommen unterschiedlich häufig in der DNA vor ($\rightarrow$ siehe Tabelle)

Modellabhängige Verfahren

- Genvorhersage auf Basis eines zu Grunde liegenden stochastischen Modells
- dieses beschreibt Eigenschaften der DNA der untersuchten Spezies

Codon Preference

- verwendet das ungleich häufige Auftreten synonymer Codons für die Genvorhersage

Codon Preference

- verwendet das ungleich häufige Auftreten synonymer Codons für die Genvorhersage
- relative Häufigkeit von Codon c bezüglich aller dazu synonymen Codons c':

$$F_R(C) = F(C) / \sum_{c' \equiv c} F(c')$$

Codon Preference

- verwendet das ungleich häufige Auftreten synonymer Codons für die Genvorhersage
- relative Häufigkeit von Codon c bezüglich aller dazu synonymen Codons c':

$$F_R(C) = F(C) / \sum_{c' \equiv c} F(c')$$

- 1.Fall: Sequenz S, gelesen in Leserahmen i, ist codierend. Wahrscheinlichkeit von S:

$$P_R^i(S) = P_R(C^i) = F_R(C_1^i) \cdot F_R(C_2^i) \cdot \ldots \cdot F_R(C_m^i)$$

Codon Preference

- verwendet das ungleich häufige Auftreten synonymer Codons für die Genvorhersage
- relative Häufigkeit von Codon c bezüglich aller dazu synonymen Codons c':

$$F_R(C) = F(C) / \sum_{c' \equiv c} F(c')$$

- 1.Fall: Sequenz S, gelesen in Leserahmen i, ist codierend. Wahrscheinlichkeit von S:

$$P_R^i(S) = P_R(C^i) = F_R(C_1^i) \cdot F_R(C_2^i) \cdot \ldots \cdot F_R(C_m^i)$$

- 2.Fall: Sequenz S, gelesen in Leserahmen i, ist nicht codierend. Annahme: alle synonymen Codons gleich häufig, also Wahrscheinlichkeit $F_{R0}(c)$ gleich $1/n_c$, wobei n_c die Anzahl der Codons, die synonym zu c sind. $\rightarrow P_{R0}(S)$

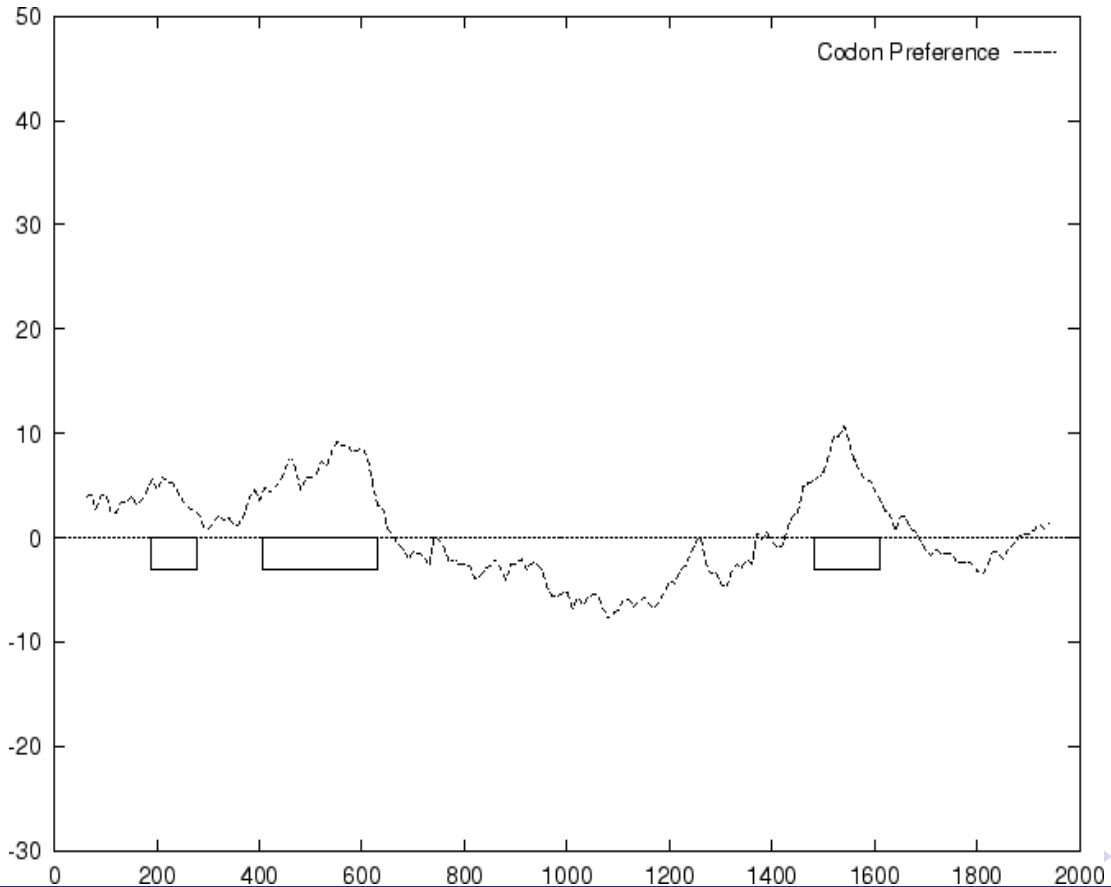
- log-likelihood rate

$$LP^i(S) = \log(P_R^i(S)/P_{R0}(S))$$

- log-likelihood rate

$$LP^i(S) = \log(P_R^i(S)/P_{R0}(S))$$

- Berechnung der LP's für überlappende Teilbereiche der Folge für alle 3 Leserahmen (Verschiebefenster)



Amino Acid Usage

- verwendet das ungleich häufige Auftreten von Aminosäuren
- beobachteter Wert wird in Relation gesetzt zu erwartetem Wert:

$$F_A(C) = \sum_{c' \equiv c} F(c')$$

Amino Acid Usage

- verwendet das ungleich häufige Auftreten von Aminosäuren
- beobachteter Wert wird in Relation gesetzt zu erwartetem Wert:

$$F_A(C) = \sum_{c' \equiv c} F(c')$$

- 1.Fall: Sequenz S, gelesen in Leserahmen i, ist codierend. Wahrscheinlichkeit von S:

$$P_A^i(S) = P_A(C^i) = F_A(C_1^i) \cdot F_A(C_2^i) \cdot \ldots \cdot F_A(C_m^i)$$

Amino Acid Usage

- verwendet das ungleich häufige Auftreten von Aminosäuren
- beobachteter Wert wird in Relation gesetzt zu erwartetem Wert:

$$F_A(C) = \sum_{c' \equiv c} F(c')$$

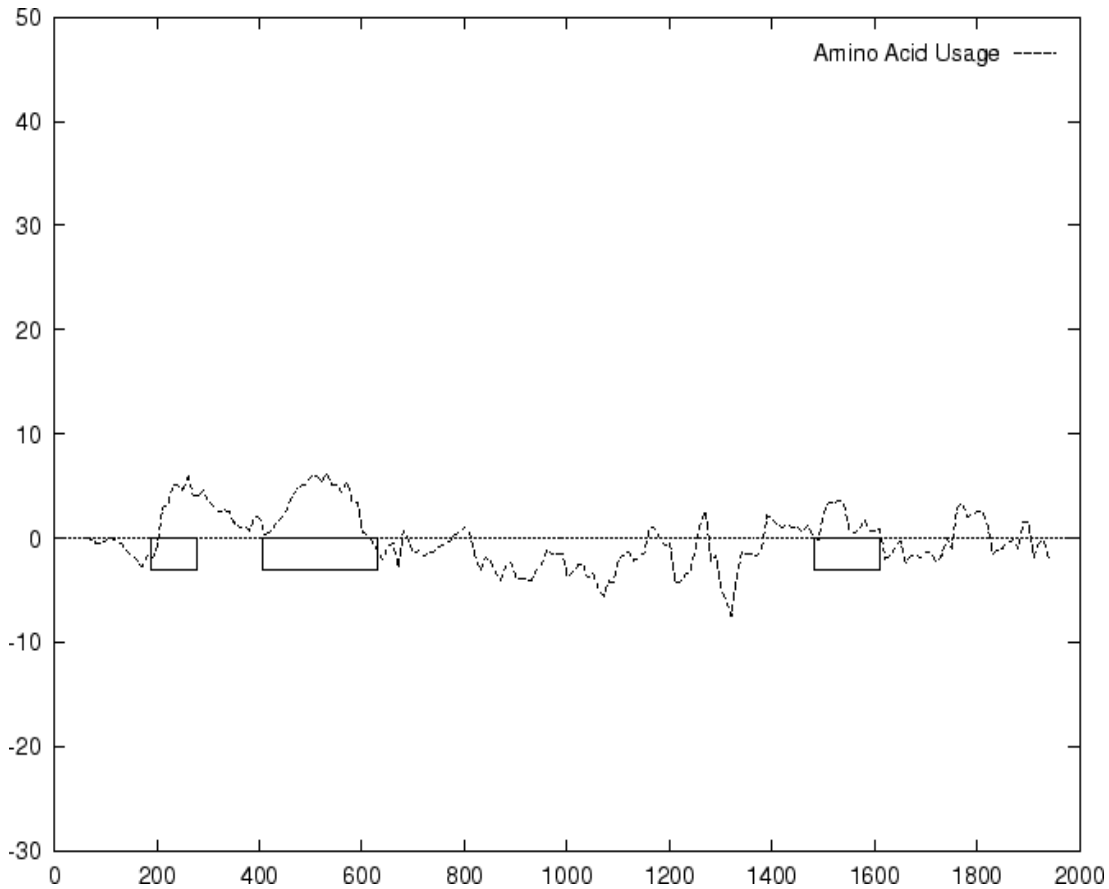
- 1.Fall: Sequenz S, gelesen in Leserahmen i, ist codierend. Wahrscheinlichkeit von S:

$$P_A^i(S) = P_A(C^i) = F_A(C_1^i) \cdot F_A(C_2^i) \cdot \ldots \cdot F_A(C_m^i)$$

- 2.Fall: Sequenz S, gelesen in Leserahmen i, ist nicht codierend. Annahme: Wahrscheinlichket einer Aminosäure proportional zur Anzahl der sie codierenden Codons, also $F_{A0}(c)$ gleich $n_c/64$, wobei n_c die Anzahl der Codons, die synonym zu c sind. $\rightarrow P_{A0}(S)$

- log-likelihood rate

$$LP^i(S) = \log(P^i_A(S)/P_{A0}(S))$$



Codon Usage

- verwendet das ungleich häufige Auftreten von Codons

Codon Usage

- verwendet das ungleich häufige Auftreten von Codons
- 1.Fall: Sequenz S, gelesen in Leserahmen i, ist codierend. Wahrscheinlichkeit von S:

$$P^i(S) = P(C^i) = F(C_1^i) \cdot F(C_2^i) \cdot \ldots \cdot F(C_m^i)$$

Codon Usage

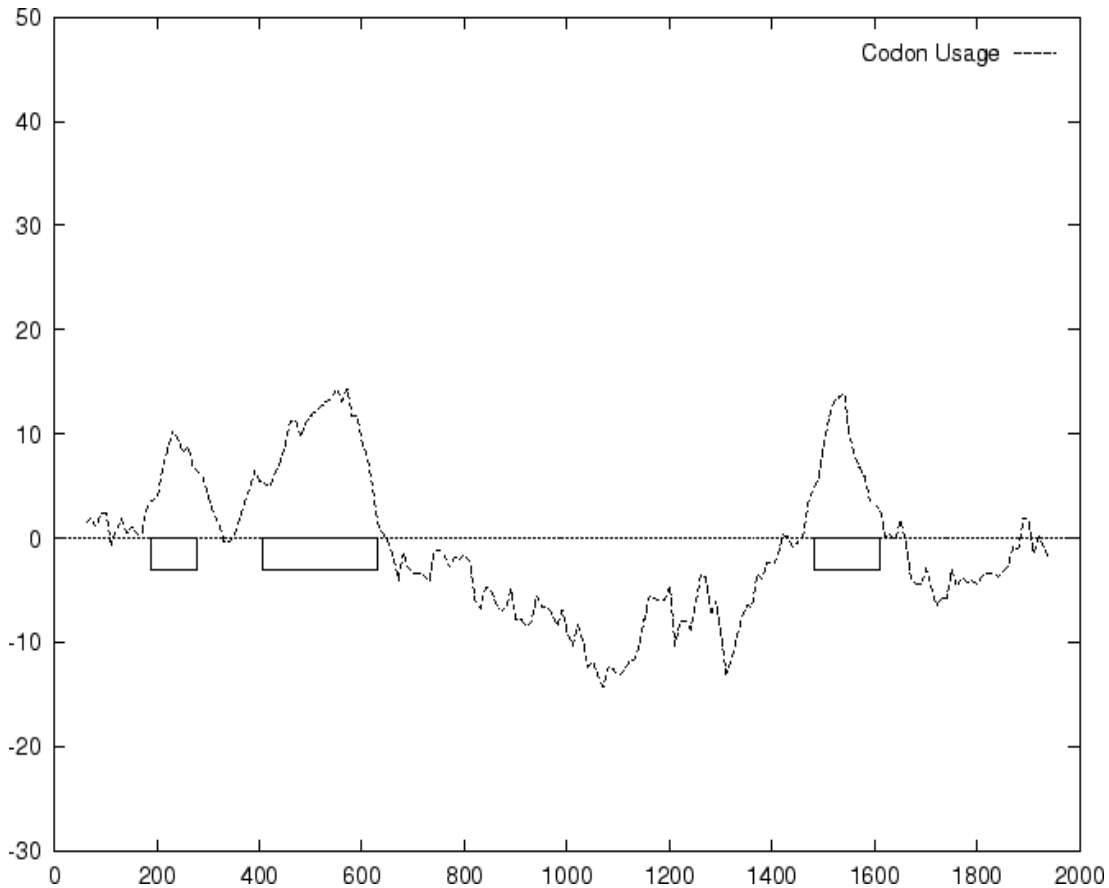
- verwendet das ungleich häufige Auftreten von Codons
- 1.Fall: Sequenz S, gelesen in Leserahmen i, ist codierend. Wahrscheinlichkeit von S:

$$P^i(S) = P(C^i) = F(C_1^i) \cdot F(C_2^i) \cdot \ldots \cdot F(C_m^i)$$

- 2.Fall: Sequenz S, gelesen in Leserahmen i, ist nicht codierend. Annahme: alle Codons gleich wahrscheinlich, also $F_0(c) = 1/64$. $\rightarrow P_0(S)$

- log-likelihood rate

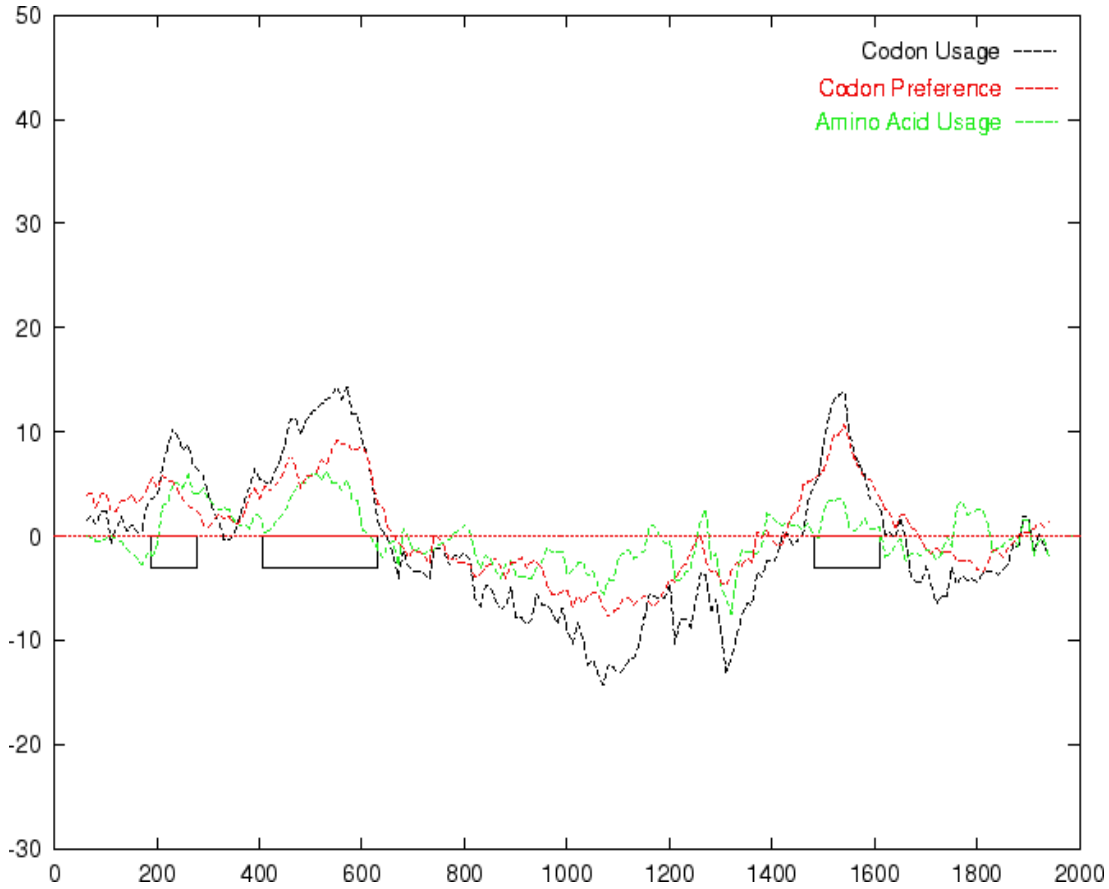
$$LP^i(S) = \log(P^i(S)/P_0(S))$$



Zusammenfassung der bisherigen Verfahren

- Untersucht wird das ungleich häufige Auftreten von
  - synonymen Codons (Codon Preference)
  - Aminosäuren (Amino Acid Usage)
  - Codons (Codon Usage)

- Codon Usage als Summe von Codon Preference und Amino Acid Usage

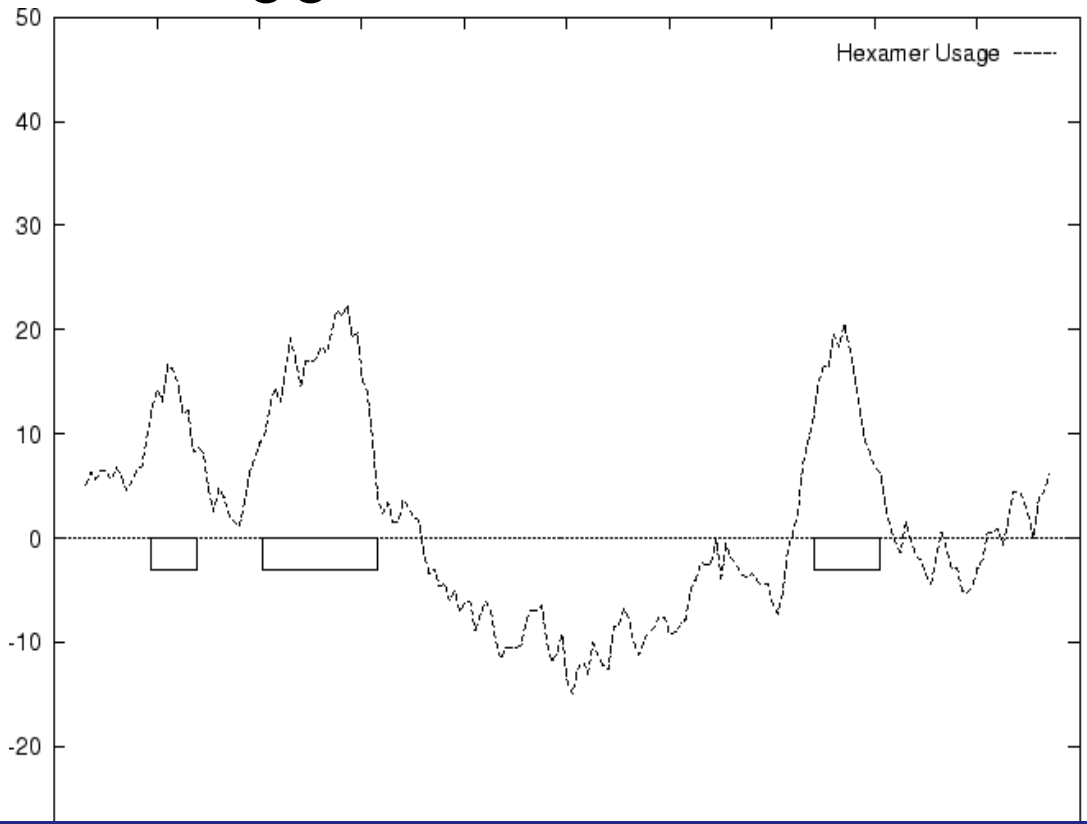


Hexamer Usage

- untersucht die Häufigkeiten von Hexameren anstelle von Codons
- daher: sechs Leserahmen notwendig
- bezieht also Abhängigkeiten zwischen Codons mit ein

Hexamer Usage

- untersucht die Häufigkeiten von Hexameren anstelle von Codons
- daher: sechs Leserahmen notwendig
- bezieht also Abhängigkeiten zwischen Codons mit ein

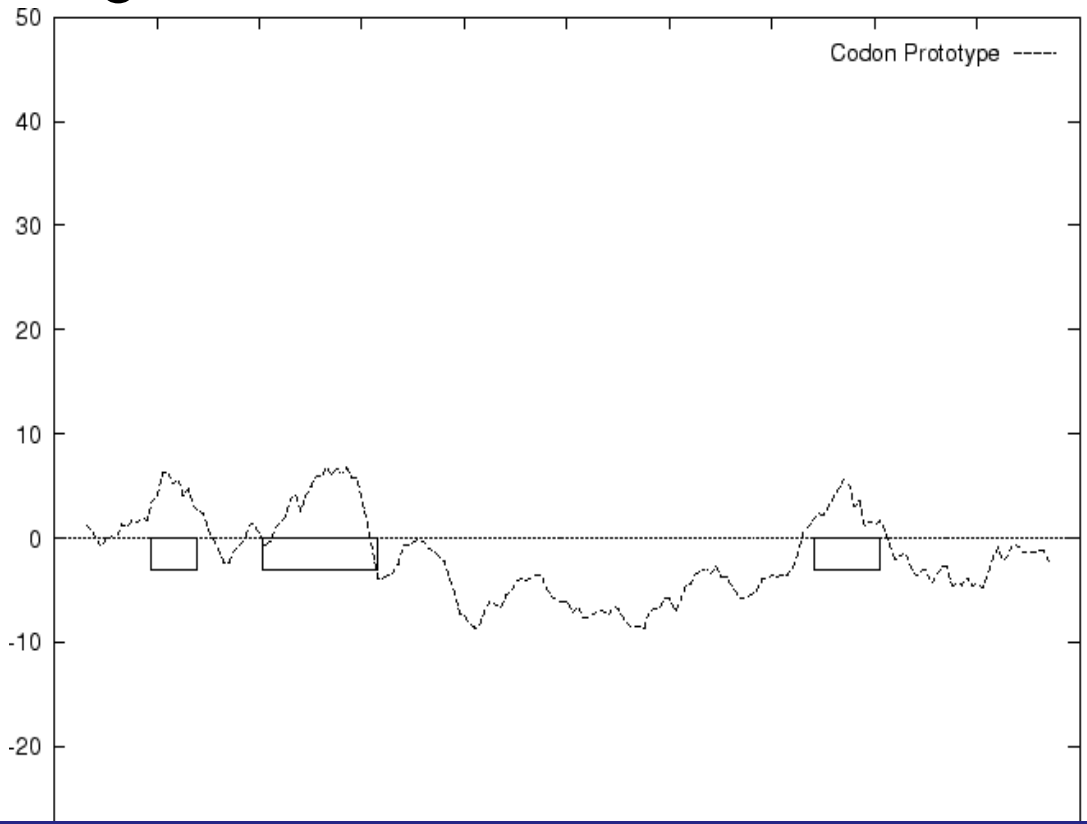


Codon Prototype

- untersucht die ungleiche Verteilung der vier Nukleotide an den drei Codonpositionen
- Grund dafür: unterschiedliche Häufigkeiten von Codons und Struktur des genetischen Codes

Codon Prototype

- untersucht die ungleiche Verteilung der vier Nukleotide an den drei Codonpositionen
- Grund dafür: unterschiedliche Häufigkeiten von Codons und Struktur des genetischen Codes

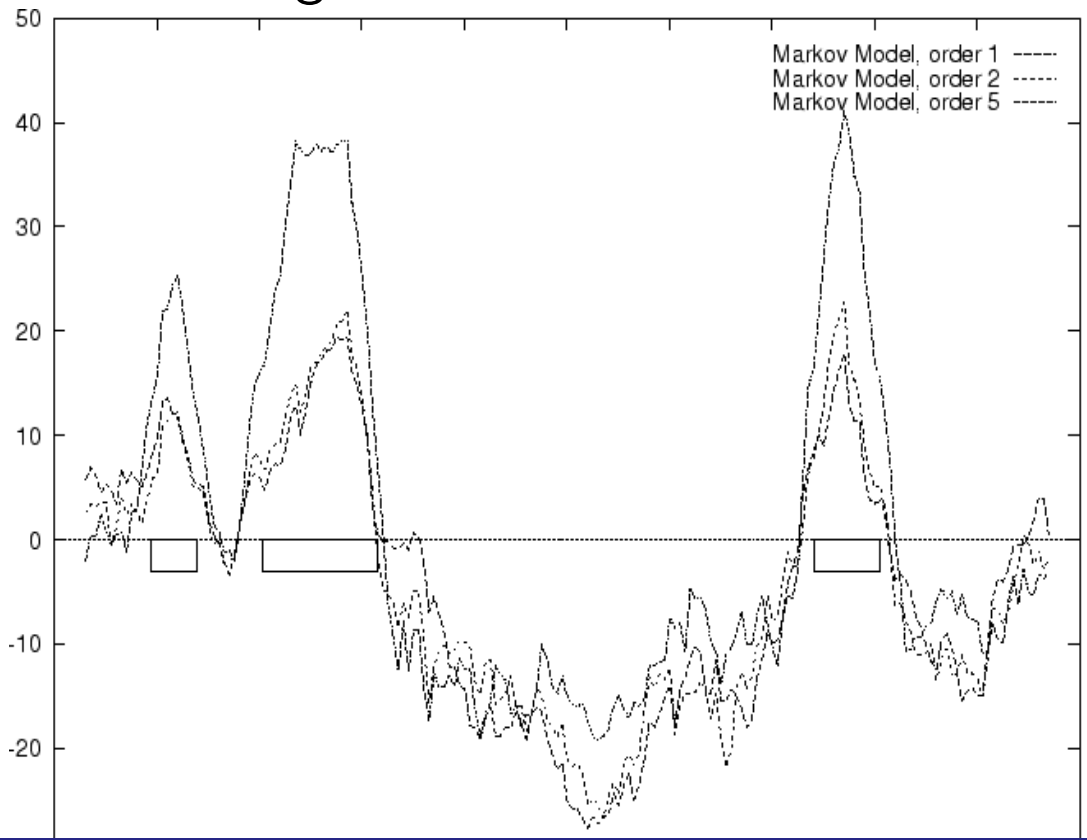


Markov Modelle

- untersucht die Wahrscheinlichkeiten von Nukleotiden an den drei Codonpositionen
- bezieht aber Vorgänger mit ein
- verschiedene Ordnungen von Markov-Modellen

Markov Modelle

- untersucht die Wahrscheinlichkeiten von Nukleotiden an den drei Codonpositionen
- bezieht aber Vorgänger mit ein
- verschiedene Ordnungen von Markov-Modellen



Modellunabhängige Verfahren

- benötigen kein zu Grunde liegendes Modell
- Vorteil: Eigenschaften der DNA der untersuchten Spezies müssen nicht bekannt sein
- Nachteil: Qualität der Ergebnisse deutlich schlechter als bei modellabhängigen Verfahren

Fourier-Analyse

- untersucht periodische Korrelationen in DNA-Sequenzen
- charakteristische Periode von 3 in codierenden Regionen
  $\longrightarrow$ hoher Wert im Spektrum für f $= 1/3$

Fourier-Analyse

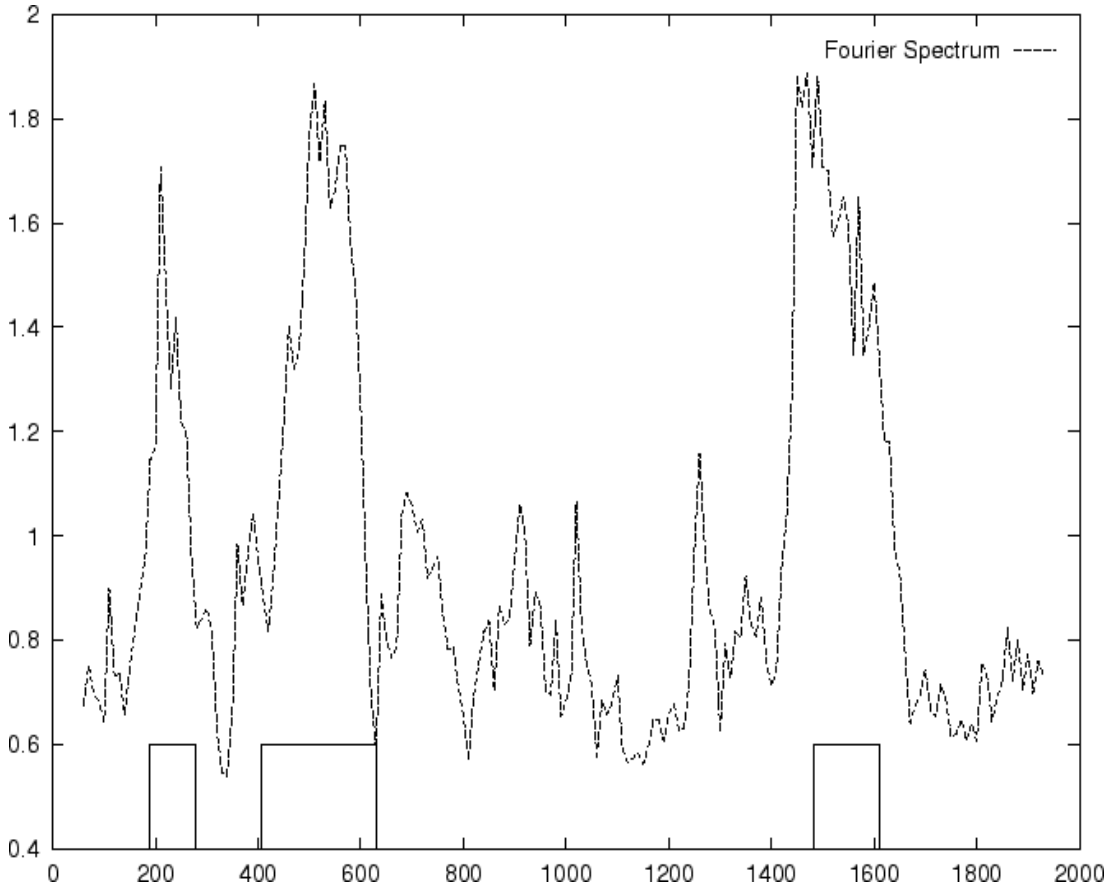
- untersucht periodische Korrelationen in DNA-Sequenzen
- charakteristische Periode von 3 in codierenden Regionen
  $\longrightarrow$ hoher Wert im Spektrum für f = 1/3
- Partielles Spektrum bezüglich Nukleotid b von DNA-Sequenz S der Länge l:

$$S_b(f) = 1/l^2 \cdot (\sum_{j=1}^{l} U_b(S_j) \cdot e^{2\pi ifj})^2$$

wobei $U_b(S_j) = 1$ falls S_j = b und 0 sonst, f diskrete Frequenz f = k/l mit k = 1,2,...,l/2
Komplettes Spektrum als Summe der vier Einzelspektren:

$$S(f) = \sum_{b \in \{A,C,G,T\}} S_b(f)$$

- Ergebnis der Analyse für f = 1/3



Vorgestellte Verfahren

Modellabhängige Verfahren

- Codon Preference
- Amino Acid Usage
- Codon Usage
- Hexamer Usage
- Codon Prototype
- Markov-Modelle

Modellunabhängige Verfahren

- Fourieranalyse

In der Praxis

- häufig Einsatz von Markov Modellen oder Kombination der vorgestellten Verfahren

In der Praxis

- häufig Einsatz von Markov Modellen oder Kombination der vorgestellten Verfahren
- Ergebnisse der vorgestellten Verfahren oft stark korrelliert

In der Praxis

- häufig Einsatz von Markov Modellen oder Kombination der vorgestellten Verfahren
- Ergebnisse der vorgestellten Verfahren oft stark korrelliert
- Sehr gute Ergebnisse: Codon Preference + Amino Acid Usage und Hexamer Usage

Quelle

- Roderic Guigó. DNA Composition, Codon Usage and Exon Prediction. *Informàtica Mèdica, Institut Municipal d'Investigació Mèdica and Department d'Estadística, Universitat de Barcelona,* Barcelona, Spain, May 1998. (incl. aller Bilder)