



האוניברסיטה
העברית
בירושלים
THE HEBREW
UNIVERSITY
OF JERUSALEM

אלגוריתמים בביו' חישובית

76558

מודלים מרקוביים
ושדראות מרקוב חבויות

תומי קפלן
16/1/2024

מודלים מרקוביים חבויים

- מודל הסתברותי/מתימטי, שמתאר תהליך סדרתי אקראי וחסר זכרון על פני מערכת מצבים חבויה.

- נתחיל בשרשרת מרקוב - תהליך מרקובי פשוט, ללא מצבים חבויים.

- המטרה: נתונה סדרת משתנים מקריים X_1, X_2, \dots נרצה למדל את ההתפלגות המשותפת שלהם בצורה קומפקטית

תכונת המרקוביות

• ניזכר בכלל השרשרת:

$$P(X_1, X_2, X_3, \dots, X_n) = P(X_1) \cdot P(X_2 | X_1) \cdot P(X_3 | X_1, X_2) \dots P(X_n | X_1, X_2, X_3, \dots, X_{n-1})$$

$$P(X_1, X_2, X_3, \dots, X_n) = P(X_1) \cdot \prod_{i=2}^n P(X_i | X_1, \dots, X_{i-1}) \quad \forall n \bullet$$

• נגדיר את תכונת המרקוביות: אי-תלות מותנית
[כלומר "היעדר זכרון"]

$$\forall_{i,j} P(X_{i+1}, \dots, X_j | X_1, \dots, X_i) = P(X_{i+1}, \dots, X_j | X_i)$$

תכונת המרקוביות

- כלומר, תת-הסדרה X_{i+1}, \dots, X_j אינה תלויה בערכי X_1, \dots, X_{i-1} בהנתן X_i

$$X_{i+1}, \dots, X_j \perp\!\!\!\perp X_1, \dots, X_{i-1} \mid X_i$$

- ובפרט $X_{i+1} \perp\!\!\!\perp X_{i-1} \mid X_i$

- ותחת הנחת המרקוביות, נוכל לשכתב

$$P(X_1, X_2, X_3, \dots, X_n) = P(X_1) \cdot P(X_2 \mid X_1) \cdot P(X_3 \mid \cancel{X_1}, X_2) \dots P(X_n \mid \cancel{X_1}, \cancel{X_2}, \cancel{X_3}, \dots, X_{n-1})$$

תכונת המרקוביות

- כלומר, תת-הסדרה X_{i+1}, \dots, X_j אינה תלויה בערכי X_1, \dots, X_{i-1} בהנתן X_i

$$X_{i+1}, \dots, X_j \perp\!\!\!\perp X_1, \dots, X_{i-1} \mid X_i$$

- ובפרט $X_{i+1} \perp\!\!\!\perp X_{i-1} \mid X_i$

- ותחת הנחת המרקוביות, נוכל לשכתב

$$P(X_1, X_2, X_3, \dots, X_n) = P(X_1) \cdot \prod_{i=2}^n P(X_i \mid X_{i-1})$$

תהליך מרקובי

- בתהליך מרקובי, סדרת המשתנים $\{X_1, \dots, X_n\}$ שייכים לא"ב סופי של מצבים $\forall_t X_t \in \{S_1, \dots, S_k\}$

- נמדל את התהליך בעזרת:

- וקטור ההתפלגות ההתחלתית $X_1 \sim \vec{P}_0$

- מטריצת מעברים $[P_t]_{i,j} = P(X_t = S_j | X_{t-1} = S_i)$

- בד"כ נניח שהתהליך הומוגני בזמן

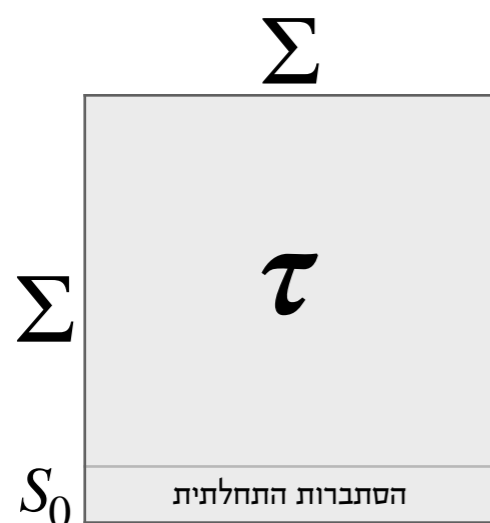
$$\forall_{i,j,t_1,t_2} P(X_{t_1} = S_j | X_{t_1-1} = S_i) = P(X_{t_2} = S_j | X_{t_2-1} = S_i)$$

מטריצת הסתברויות מעבר

• נגדיר: $\tau_{a,b} \triangleq P(X_t = b | X_{t-1} = a) = P(a \rightarrow b)$

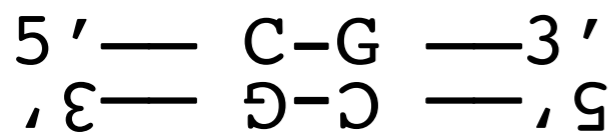
• וכמובן $\forall_{a,b} \tau_{a,b} \geq 0$ $\forall_a \sum_b \tau_{a,b} = 1$

- לפעמים שימושי להוסיף מצב התחלה $S_0 \notin \Sigma$ [פיקטיבי] ולהגדיר את וקטור ההתפלגות ההתחלתית כשורה נוספת ב- τ



דוגמא: איי CpG

- בגנום, כאשר ציטוזין מופיע לפני גואנין, ב-80% מהמקרים הציטוזין יעבור מתילציה

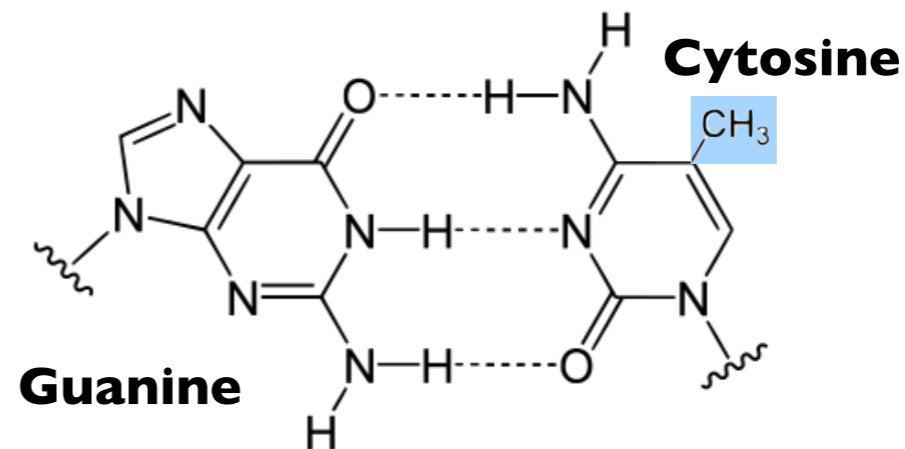
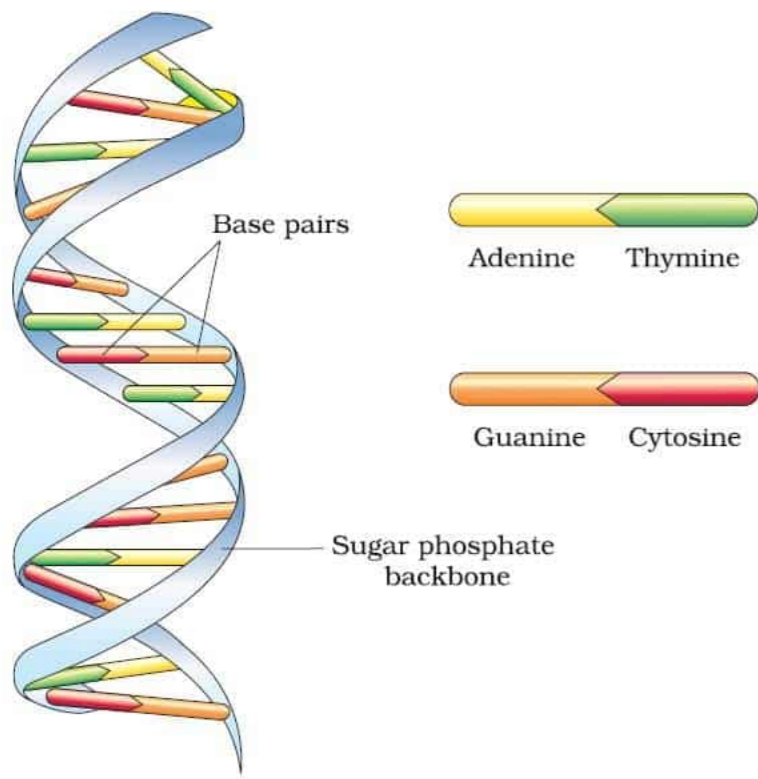


ל-5-מתיל-ציטוזין

- המתילציה תישמר לאורך כל חי התא, ותגדיר

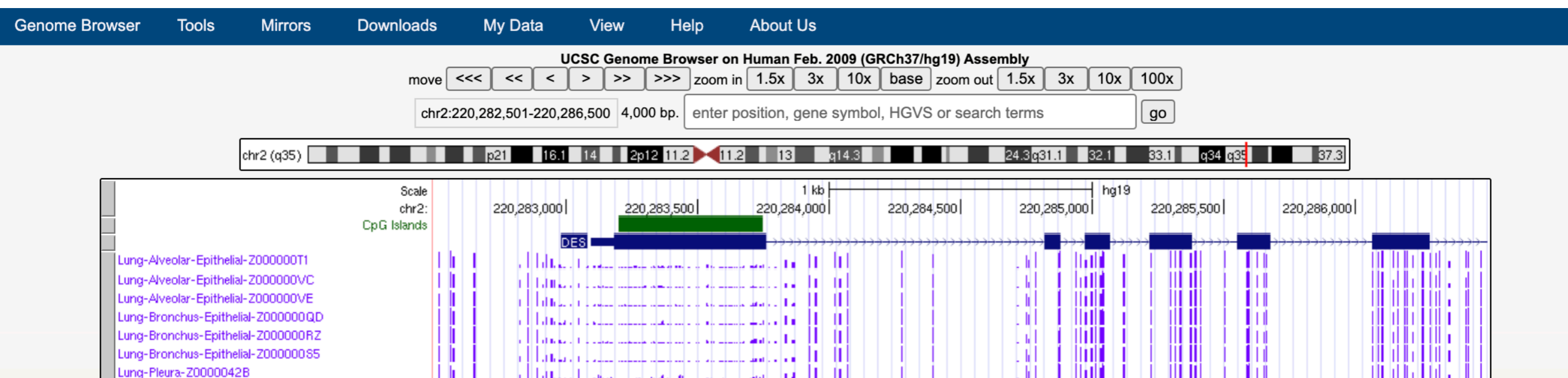
את ה"זהות התאית" שלו ואיזה גנים הוא

יבטא או ישתיק



דוגמא: איי CpG

- איי CpG מוגדרים כ:
 - איזורים גנומיים קצרים (200 בסיסים ומעלה)
 - בעלי אחוז G/C גדול מ-50%
 - בעלי אחוז CpG גדול מ-2.6%
- [בשאר הגנום, שכיחות CpG קטנה מ-1%]



CpG י"ן :NλλIT

▶chr2:220282501-220286500

GAGGCTCAGGGCTAGCTCGCCCATAGACATACATGGCAGGCAGGCTTTGGCCAGGATCCCTCCGCCTGCCAGGCGTCTCCCTGCCCTCCCTTCCTGCCTA
GAGACCCCCACCCTCAAGCCTGGCTGGTCTTTGCCTGAGACCCAAACCTCTTCGACTTCAAGAGAATATTTAGGAACAAGGTGGTTTAGGGCCTTTCCCTG
GGAACAGGCCTTGACCCTTTAAGAAATGACCCAAAGTCTCTCCTTGACC AAAAGGGGACCCTCAA ACTAAAGGGAAGCCTCTCTTCTGCTGTCTCCCT
GACCCCACTCCCCCACCACAGGACGAGGAGATAACCAGGGCTGAAAGAGGCCCGCCTGGGGGCTGCAGACATGCTTGCTGCCCTGGCGAAGGAT
TGGCAGGCTTGCCCGTCACAGGACCCCGCTGGCTGACTCAGGGGCGCAGGCCTCTTGCGGGGGAGCTGGCCTCCCCGCCCCACGGCCACGGGCCGCC
TTTCCTGGCAGGACAGCGGGATCTTGACAGCTGTCAGGGGAGGGGAGGCGGGGGCTGATGTCAGGAGGGATACAAATAGTGCCGACGGCTGGGGGCCCTGT
CTCCCTCGCCGCATCCACTCTCCGGCCGGCCGCCTGCCCGCCGCCTCCTCCGTGCGCCCGCCAGCCTCGCCCGCGCCGTACCATGAGCCAGGCCTACT
CGTCCAGCCAGCGCGTGTCTCCTACCGCCGCACCTTCGGCGGGGCCCGGGCTTCCCACTCGGCTCCCCGCTGAGTTCGCCCGTGTTCGCCGCGGGCGGG
TTTCGGCTCTAAGGGCTCCTCCAGCTCGGTGACGTCCCGCGTGTACCAGGTGTCGCGCACGTGGGGCGGGGCCGGGGCCTGGGGTCTGCTGCGGGCCAGC
CGGCTGGGGACCACCCGCACGCCCTCCTCCTACGGCGCAGGCGAGCTGCTGGACTTCTCACTGGCCGACGCGGTGAACCAGGAGTTTCTGACCACGCGCA
CCAACGAGAAGGTGGAGCTGCAGGAGCTCAATGACCGCTTCGCCAACTACATCGAGAAGGTGCGCTTCCTGGAGCAGCAGAACGCGGCCTCGCCGCCGA
AGTGAACCGGCTCAAGGGCCGCGAGCCGACGCGAGTGGCCGAGCTCTACGAGGAGGAGCTGCGGGAGCTGCGGGCGCCAGGTGGAGGTGCTCACTAACCCAG
CGCGCGCGCTCGACGTCGAGCGCGACAACCTGCTCGACGACCTGCAGCGGCTCAAGGCCAAGTGAGGGCCCGGCACCCCACTCCTCTTTCTGCGGGC
AGGGCACAGGAGGCTAGGCCTGGGGTCTGGGGTCCCGCTGTCAGCACCTGCCTTCTCCCGGGGCCCGGGACCCTCCTCCTGCCCCATGTGGAGAAAGGGTC
CTCCACCTGTGTGTTTCAAGGGGCCGTGACCTCCAGGTCTCTCCCTGCGATCCCATCTTGACACAGGAGTTTTCTTGGGGACATAGATCAGGGGGTGA
TATGGGAGAATTTAGGGGACCCGGTGCCTGTGGACAGCCCGTTAAAAGCATTTTAAAGATGCTGGGGCGATATTTATGGGGTCAGGTAGTTGATGGGC
AGAGGAAGGGCTGCAGGAGGCCAGAGGGCAGTGTAGCCAGAGGGAGAAGGGAGGCTGATAGGAGACAGGGAAAGCAGGGCAAGGGCCAGAGTCCAAGC
AACAGCTCTCAGCTCAGCTGTGATGAGGCCCTGGGGGAGGTGGGGGGAGGGGGGAGCTTGGCCCTGGGGCCTTGCCGAGACTGTGTCTTTTTTACAAGGTG
AATGGACAGGCTGGAGAAAAGGGAGTAGGTGGGGGTACAGCTCTCAGAGAGCTTGGGAGGACCTGACTGTAGACTTCACCAGGCTCCAAGAACGAAA
GGGCAGCAAGTGTAGCATATTTGTTGGTCCCACTTCTGACAGGCCAAGTGAGCACAGTCACCCTCCTGCCACCAAAGTCATAAATATTAATTGAGCAGCT
ATATTGGCCAGGCTGGAGCTGGGAACCAGAAACACAGAGGTGGATAAAATAGACACAGTTTTCTAACCCCAAGGGAGGTACACAGTCTGGTGGGGACATAG
ACTTCAAGGGTGTGGCTCCTGGGCAGAGATTGGGCCACTTCCCTGTGCCCTCCCTGGGTGGGTGGGGCCTCTCCACTCCCTGTCTCTCCTGCCTCTACCCA
GCAGCCAGGCCCTCCCGCTCTGTCTGGACCCACCCCTGGTCAGCCCCCGCCAGTCGTTTCCACTGCCAGCTTTATCACCCGCAACTGTCTGTCTTTC
TGTCTGTCCACCCAGGCTGCAGGAGGAGATTGAGTTGAAGGAAGAAGCAGAGAACAATTTGGCTGCCTTCCGAGCGGTGAGTGCCCTTCTTTTCCCCTT
GCATGGCCTCTGGCCTTGCTCTGCCCCACCTGGGTGGCGGTGACCATGTCCTTCTCGCTTGGCCTCTCCAGGACGTGGATGCAGCTACTCTAGCTCGCA
TTGACCTGGAGCGCAGAATTGAATCTCTCAACGAGGAGATCGCGTTCCCTAAGAAAGTGCAATGAAGAGGTATACCTTGGCCCTCCTCCTGGGGTCACTG
GGCCATGGGGAAAGCAGCCGGAAGTGGGGTGGGGTGGGCTCTGGCTGGGAATAGGGGTGTGAGGGTGTGTGGGGCCCTGAGAGGGGACTGAAGCC
CAGTCATGCCCTACAGGAGATCCGTGAGTTGCAGGCTCAGCTTCAGGAACAGCAGGTCCAGGTGGAGATGGACATGTCTAAGCCAGACCTCACTGCCGCC
CTCAGGGACATCCGGGCTCAGTATGAGACCATCGCGGCTAAGAACAATTTCTGAAGCTGAGGAGTGGTACAAGTCGAAGGTGGGTGGCCTCGCCCGGGGAC
TGGCATCTCCGTCCCCCTGAATCCAGCTTGGATGTGCTGCCTGTGGTACCATCCATGGGAGGAGAGCCAGAGGCTTCATGCTCCCTTGCTCATCCCTA
CCCGTGCCCTGCATCCTTCTCATTTTTTGGGCCCTTTCTCTGCCCTTAGGTGTGACACTGACCCAGGCAGCCAACAAGAACAACGACGCCCTGCGCCAG
GCCAAGCAGGAGATGATGGAATACCGACACCAGATCCAGTCCCTACACCTGCGAGATTGACGCCCTGAAGGGCACTGTGAGTCCCTGCCACCTGGCCAGG
CCCTGCCCTTCTGTCTGCAGTTCACACCCTCACTTTGTGACCTTGGGCCCATCATAGATCCTCTCTGGGCCTTCATCTACTTAAATCTACAATAGGGG
TAAAACCAGACAAGTGGATTCCAGTTGGATGCTAAGGAATCAGGGGTTCCCTGGGCATCTACCTATGTGGGGACTGTGAGGCTGAATGCAATGTTCTTTG
TATCTATTTTATTCTGAGTGTTCACATATAGACTTAATTTGAGTTCAGGGTTCAACATGGCCTGGACCTGACCATCTGGAGTTGCCTGCCAGCCCAAAG
CTTTCTTTGGGCTGCTAGTGTCTCTTCCCTTCCCTTGACCTGGGTTCGCCCTCCTGCAGAACGATTCCTGATGAGGCAGATGCGGGAAATTGGAGGAC
CGATTTGCCAGTGAGGCCAGTGGCTACCAGGACAACATTCGCGCCCTGGAGGAGGAAATCCGGCACCTCAAGGATGAGATGGCCCGCCATCTGCGCGAGT
ACCAGGACCTGCTCAACGTGAAGATGGCCCTGGATGTGGAGATTGCCACCTACCGGAAGCTGCTGGAGGGAGAGGAGAGCCGGTGGAGGGGCCAGGCAGGA
GCCCCGAGTGGGAGGTGCGGGGTGCTGGGTGGTCCATTTCTGTCCCAGGAGGCTCGAGATTACTGATTACCTCAACAAGACCTGGAAACAATTTTTTTTT
TTTTTGAGATGGAGTTTCGCTCTTGTGCGCCAGTCTGGAGTGCAATGGCACCATCTTGGCTCACTGCAACCTCCGCCTCCTGGGTTCAGCAATTCCTCCT



CpG י"ן :NΛλIT

```

>chr2:220282501-220286500
GAGGC TCAGG GCTAG GCTC GC CCATAGACATACATG GCAG GCAG GCTTTG GC CAGGATCCCTCC GCCT GC CAG GC GTCTCCCT GC CCTCCCTTCCT GC CTA
GAGACCCCCACCCTCAAG GCTG GCTGGTCTTT GCCTGAGACCCAAACCTCTTCGACTTCAAGAGAATATTTAGGAACAAGGTGGTTTAGG GCCTTTCCCTG
GGAACAG GCCTTGACCCTTTAAGAAATGACCCAAAGTCTCTCCTTGACCAAAAAGGGGACCCTCAAATAAAGGGAAG GCCTCTCTTCT GC TGTCTCCCT
GACCCCACTCCCCCCACCACAGGACGAGGAGATAACCAGG GCTGAAAGAG GC CC GCCTGGGG GC TGCAGACAT GC TTTGC TGC CT GC CCTG GC GAAGGAT
TG GCAG GCTT GC CCGT CACAGGACCCCC GC TG GCTGACTCAGGG GCGCAG GCCTCTT GC GGGGGAG GC TG GCCTCCCC GC CCCACG GC CACGG GC GCGC
TTTCCTGG CAGGACAG GGGATCTT GCAG GCTGTCAGGGGAGGGGAG GC GGGG GC TGATGTCAGGAGGGATACAAATAGT GC CGACG GC TGGGG GC CCTGT
CTCCCTC GC C GCATCCACTCTCCG GC CG GC GCCT GC CC GC GCCTCCTCCGT GCGC CC GC CAG GC CTC GC CC GCGC CGTCACCATGAG GC CAG GCCTACT
CGTCCAG GC CAG GCGC GTGTCCTCCTACC GC C GC ACCTTCGC GGG GC CCCGG GC TCCCACTCG GC TCCCC GC TGAGTTC GC CCGTGTTC GC GGG GC GGG
TTTCG GC TCTAAGG GC TCTCCA GC TCGGTGACGTCCC GC GTGTACCAGGTGTC GCGC ACGTCCG GC GGG GC CGGGG GC CTGGGGT GC TGC GGG GC CAG GC
CG GC TGGGGACACC GC AC GC CCTCCTCCTACG GCGCAG GC GAG GC TGC TGGACTTCTCACTG GC CGAC GC GGTGAACCAGGAGTTTCTGACCAC GCGCA
CCAACGAGAAGGTGGAG GC TGCAGGAG GC TCAATGACC GC TTC GC CAACTACATCGAGAAGGT GCGC TTCCTGGAG GC CAGCAGAAC GC GCGC GC TGC GCGC GA
AGTGAACCG GC TCAAGG GC CGC GAG GC CGAC GC GAGTGC CGAG GC TCTACGAGGAGGAG GC TGC GGGAG GC TGC GCGC CAGGTGGAGGT GC TCACTAACCAG
C GCGC GCGC GTCGACGTGAG GCGC GACAACCT GC TCGACGACCT GC CAG GC GGC TCAAG GC CAAGTGAAG GC CCG GC ACCCCAGACTCCTCTTTCT GC GGG GC
AGG GC ACAGGAG GC TAG GC CTGGGGTCTGGGGTCCC GC TGTCA GC ACCT GCCTTCTCCCGGG GC CCGGACCCTCTCCT GC CCCATGTGGAGAAAGGGTC
CTCCACCTGTGTGTTTCAAGGG GC CGTGACCTCAGGTCTCTCCCCCT GC GATCCCATCTT GC ACAGGAGTTTTCTTGGGGACATAGATCAGGGGGTGA
TATGGGAGAATTTAGGGGACCCGGT GC CCTGTGGACAG GC CCCGTTAAAA GC ATTTTAAAGAT GC TGGG GC GATATTTATGGGGTCAAGGTAGTTGATGG GC
AGAGGAAGG GC TGCAGGAG GC CCAGAGG GC AGTGTAG GC CAGAGGGAGAAGGGAG GC TGATAGGAGACAGGGAAAG GC AGG GC AAGG GC CCAGAGTCCAAG GC
AACAG GC TCTCAG GC TCA GC TGTGATGAG GC CCTGGGGGAGGTGGGGGGAGGGGGGAG GC TTG GC CCTGGG GC CTT GC CGAGACTGTGTCTTTTTTACAAGGTG
AATGGACAG GC TGGAGAAAAGGGAGTAGGTGGGGGT CACAG GC TCTCAGAGAG GC TTGGGAGGACCTGACTGTAGACTTCACCAG GC TCCAAGAACGAAA
GG GC AGC AAGTGTAG GC ATATTTGTTGGTCCCACCTTCTGACAG GC CAAGTGA GC ACAGTCACCCTCCT GC CACCAAAGTCATAAATATTAATTGAG GC AGC T
ATATTGGC CAG GC TGGAG GC TGGGAACCAGAAACACAGAGGTGGATAAAATAGACACAGTTTTCTAACCCAGGGAGGTCACACAGTCTGGTGGGGACATAG
ACTTCAAGGGTGTG GC TCTGG GC AGAGATTGG GC CACTTCCCTGT GC CCTCCCTGGGTGGGTGGG GC CTCTCCACTCCCTGTCTCTCCT GC CTCTACCCA
GC AGC CAG GC CCTCCC GC TCTGTCTGGACCCACCCCTGGTCA GC CCCC GC CAGTCGTTTCCACT GC CAG GC TTTATCACCC GC AACTGTCTGTCTTTC
TGTCTGTCCCACCCAG GC TGCAGGAGGAGATTCAAGTTGAAGGAAGAA GC CAGAGACAATTTG GC TGC TTTCCGAG GC GGTGAGT GC CCTTCTTTTCCCCTT
GC ATG GC CTCTG GC CTT GC TCT GC CCCACCTGGGTG GC GGTGACCATGTCCTTCTC GC TTG GC CTCTCCAGGACGTGGAT GC AGC TACTCTAG GC TC GC A
TTGACCTGGAG GC GC AGAATTGAATCTCTCAACGAGGAGATC GC GTTCCCTTAAGAAAGT GC ATGAAGAGGTATACCTTG GC CCTCTTCCTGGGGTCACTG
GGC CATGGGGAAAG GC AGC CGGAAAGTGGGGTTGGGGTGAG GC TCTG GC TGGGAATAGGGGTGTGAGGGT GC TGTGTGG GC CCTGAGAGGGGACTGAAG GC
CAGTCAT GC CCTACAGGAGATCCGTGAGTT GC AG GC TCAG GCTTCAGGAACAG GC CAGGTCCAGGTGGAGATGGACATGTCTAAG GC CAGACCTCACT GC CCGC
CTCAGGGACATCCGG GC TCAGTATGAGACCATC GC GGC TAAGAACATTTCTGAA GC TGAGGAGTGGTACAAGTCGAAGGTGGGTG GC CTC GC CCGGGGAC
TG GC ATCTCCGTCCCCTGAATCCA GC TTGGATGT GC TGC CTGTGGTACCATCCATGGGAGGAGA GC CCAGAG GC TTCAT GC TCCCTT GC TCATCCCTA
CCCGT GC CCT GC ATCCTTCTCATT TTTTGG GC CCTTTTCTCT GC CCTTAGGTGT CAGACCTGACCCAG GC CAG GC CAACAAGAACAACGAC GC CCT GC GCGC CAG
GC CAAG GC AGGAGATGATGGAATACCGACACCAGATCCAGTCCCTACACCT GC GAGATTGAC GC CCTGAAGG GC ACTGTGAGTCCCT GC CCACCTG GC CAGG
CCCT GC CCTTCTCTGTCT GC AGTTCACACCCTCACTTTGTGACCTTGG GC CCATCATAGATCCTCTCTGG GC CTTTCTACTTAAATCTACAATAGGGG
TAAAACCAGACAAGTGGATTCCAGTTGGAT GC TAAGGAATCAGGGGTTCCTGG GC ATCTACCTATGTGGGGACTGTGAG GC TGAAT GC AATGTTCTCTTTG
TATCTATTTTATTCTGAGTGTTCACATATAGACTTAATTTGAGTTCAGGGTTCAACATG GC CTGGACCTGACCATCTGGAGTT GC CTGC CAG GC CCCAAAG
CTTTCTTTGG GC TGC TAGTGTCTCTTCCCTTCTTGACCTGGGTTCCTCCT GC CAGAACGATTCCTGATGAG GC CAGAT GC GGGAAATTGGAGGAC
CGATTT GC CAGTGAAG GC CAGTGC TACCAGGACAACATTT GCGCGC CTGGAGGAGGAAATCCG GC ACCTCAAGGATGAGATGC CC GC CATCT GC GCGC GAGT
ACCAGGACCT GC TCAACGTGAAGATGC CCTGGATGTGGAGATT GC CACCTACCGGAA GC TGC TGAGGGGAGAGGAGA GC CGGTGAAGGG GC CAG GC CAGGA
GC CCGAGTGGGAGGT GC GGGGT GC TGGGTGGTCCATTTCTGTCCCAGGAG GC TCGAGATTACTGATTACCTCAACAAGACCTGGAAACAATTTTTTTTTT
TTTTTGAGATGGAGTTTC GC TCTTGTGC GC CCAGTCTGGAGT GC AATG GC ACCATCTTG GC TCACT GC AACCTCC GC CTCTGGGTTCAG GC AATTTCTCCT

```

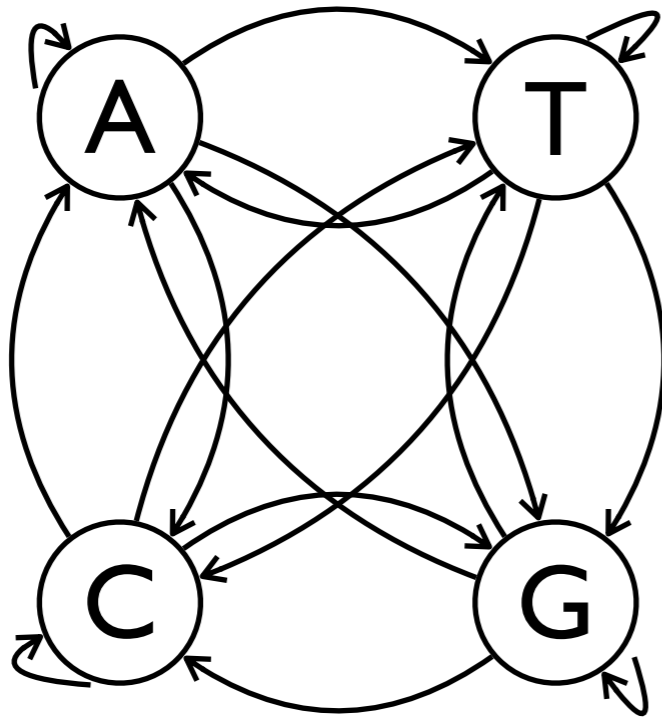
CG

CpG י"ן :נלללל

```
>chr2:220282501-220286500
GAGGCTCAGGGCTAGCTCGCCCATAGACATAACATGGCAGGCAGGCTTTGGCCAGGATCCCTCCCTGCCAGGCGTCTCCCTGCCCTCCCTTCCTGCCTA
GAGACCCCCACCCTCAAGCCTGGCTGGTCTTTGCCTGAGACCCAAACCTCTTCTCAAGAGAATATTTAGGAACAAGGTGGTTTAGGGCCTTTCCCTG
GGAACAGGCCTTGACCCTTTAAGAAATGACCCAAAGTCTCTCCTTGACCAAAAGGGGACCCTCAAATAAGGGGAAGCCTCTCTTCTGCTGTCTCCCT
GACCCCACTCCCCCACCACCAGGACGAGGAGATAACCAGGGCTGAAAGAGGCCCTGGGGGCTGCAGACATGCTTGCTGCCCTGGCGAAGGAT
TGGCAGGCTTGCCCGTCACAGGACCCCGCTGGCTGACTCAGGGCGCAGGCCTCTTGCGGGGAGCTGGCCTCCCGCCCCACGCGCCAGGGC
TTTCCCTGGCAGGACAGCGGGATCTTGCAGCTGTGAGGGGAGGGGAGGGGGCTGATGTGAGGAGGGATACAAATAGTGCAGCGGCTGGGGGCCCTGT
CTCCCTCGCCGCATCCACTCTCGGCGCGCCCTGCCCGCCCTCCTCGTGCCTCCAGCCTCGCCCGCGCGTCCACCATGAGCCAGGCCTACT
CGTCCAGCCAGCGCGTGTCTCTTACCGCCGACCTTTCGGCGGGGCCCGGGCTTCCCACTCGGCTCCCGCTGAGTTCCCGTGTTCGCGGGGCGGG
TTTTCGGCTCTAAGGGCTCCTCCAGCTCGGTGACGTCCCGCGTGTACCAGGTGTCCCGCACGTTCGGCGGGGCCTGGGGTCCGCTGCGGGCCAGC
CGGCTGGGGACCACCCGCAAGCCCTCCTCCTACGGCGCAGGCGAGCTGCTGGACTTCTCACTGGCCGACCGCGTGAACCAGGAGTTTCTGACCA
CCAAAGAGAAGGTGGAGCTGCAGGAGCTCAATGACCGCTTCGCAACTACATCGAGAAGGTGCGCTTCCTGGAGCAGCAGAAAGCGCGCGCTCGCCGCGA
AGTGAACCGGCTCAAGGGCGCGGAGCGAGCGCGAGTGGCGAGCTCTACGAGGAGGAGCTGCGGGAGCTGCGCGCCAGGTGGAGGTGCTCACTAACAG
CGCGCGCGCGTTCGACGTTCGAGCGCGACAACCTGCTCGACGACCTGCAGCGCTCAAGGCCAAGTGGAGGGCCCGGCACCCCAAGACTCCTCTTTCTGCGGGC
AGGGCACAGGAGGCTAGGCCTGGGGTCTGGGGTCCCGCTGTGAGCACCTGCCCTTCTCCCGGGCCCGGGACCCTCCTGCCCCATGTGGAGAAAGGGTC
CTCCACCTGTGTGTTTCAAGGGGCGGTGACCTCAGGTCTCTCCCTCGATCCCCTGTCACAGGAGTTTCTTTGGGGACATAGATCAGGGGGTGA
TATGGGAGAATTTAGGGGACCCGGTGCCCTGTGGACAGCCCGTTAAAAGCATTTTAAAGATGCTGGGGCGATATTTATGGGGTCAGGTAGTTGATGGG
AGAGGAAGGGCTGCAGGAGGCCAGAGGGCAGTGTAGCCAGAGGGAGAAGGGAGGCTGATAGGAGACAGGGAAAGCAGGGCAAGGGCCAGAGTCCAAGC
AACAGCTCTCAGCTCAGCTGTGATGAGGCCCTGGGGGAGGTGGGGGGAGGGGGAGCTTGGCCCTGGGGCCTTGCAGACTGTGTCTTTTTTACAAGGTG
AATGGACAGGCTGGAGAAAAGGGAGTAGGTGGGGGTACAGCTCTCAGAGAGCTTGGGAGGACCTGACTGTAGACTTCACCAGGCTCCAAGAAAGAAA
GGGCAGCAAGTGTAGCATATTTGTTGGTCCCACTTCTGACAGGCCAAGTGAGCACAGTCACCCTCCTGCCACCAAAGTCATAAATATTAATTGAGCAGCT
ATATTGGCCAGGCTGGAGCTGGGAACCAGAAACACAGAGGTGGATAAAATAGACACAGTTTCTAACCACAGGGAGGTACACAGTCTGGTGGGGACATAG
ACTTCAAGGGTGTGGCTCCTGGGCAGAGATTGGGCCACTTCCCTGTGCCCTCCCTGGGTGGGTGGGGCCTCTCCACTCCCTGTCTCTCCTGCCTCTACCCA
GCAGCCAGGCCCTCCCGCTCTGTCTGGACCACCCCTGGTCAGCCCCCGGCCAGTCGTTCACACTGCCAGCTTTATCACCAGCAACTGTCTGTCTTTC
TGTCTGTCCACCAGGCTGCAGGAGGAGATTAGTTGAAGGAAGAAGCAGAGAACAATTTGGCTGCCTTCAGCGGTGAGTGCCCTTCTTTTCCCTT
GCATGGCCTCTGGCCTTGTCTGCCCCACCTGGGTGGGTGACCATGTCCTTCTCTGGCCTCTCCAGGACCGTGGATGCAGCTACTCTAGCTCGCA
TTGACCTGGAGCGCAGAATTGAATCTCTCAAAGGAGATCGCGTTCCCTAAGAAAGTGCAATGAAGAGGTATACCTTGGCCCTCTTCCCTGGGGTCACTG
GGCCATGGGGAAAGCAGCGGAAAGTGGGGTGGGGTGGGGTGGGGTGGGGTGGGGTGGGGTGGGGTGGGGTGGGGTGGGGTGGGGTGGGGTGGGGTGGGGT
CAGTCATGCCCTACAGGAGATCGTGAGTTGCAGGCTCAGCTTCAGGAACAGCAGGTCCAGGTGGAGATGGACATGTCTAAGCCAGACCTCACTGC
CTCAGGGACATCCGGCTCAGTATGAGACCATCGCGGCTAAGAACAATTTCTGAAGCTGAGGAGTGGTACAAGTCGAAGGTGGGTGGCCTCGCCCGGGAC
TGGCATCTCCGTCCCCTGAATCCAGCTTGGATGTGCTGCCTGTGGTACCATCCATGGGAGGAGAGCCAGAGGCTTCATGCTCCCTTGTCTCATCCCTA
CCCGTGCCCTGCATCCTTCTCATTTTTGGGCCCCCTTTCTCTGCCCTTAGGTGTGACCTGACCCAGGCAGCCAACAAGAACAACGACCGCCTGCGCCAG
GCCAAGCAGGAGATGATGGAATACCGACACCAGATCCAGTCCCTACACCTGCGAGATTGACCGCCTGAAGGGCACTGTGAGTCCCTGCCACCTGGCCAGG
CCCTGCCCTTCCCTGTCTGCAGTTCACACCTCACTTTGTGACCTTGGGCCATCATAGATCCTCTCTGGGCCTTCATCTACTTAAATCTACAATAGGGG
TAAAACCAGACAAGTGGATTCCAGTTGGATGCTAAGGAATCAGGGGTTCCCTGGGCATCTACCTATGTGGGGACTGTGAGGCTGAATGCAATGTTCTTTG
TATCTATTTTATTTCTGAGTGTTCACATATAGACTTAATTTGAGTTCAGGGTTCAACATGGCCTGGACCTGACCATCTGGAGTTGCCTGCCAGCCCAAG
CTTTCTTTGGGCTGCTAGTGTCTCTCCCTTCCCTTGACCTGGGTCCCCCTCTCCTGCAGAAAGATTCCCTGATGAGGCAGATCGGGGAATTGGAGGAC
CGATTTGCCAGTGAGGCCAGTGGCTACCAGGACAACATTCGCGCCTGGAGGAGGAAATCGGCACCTCAAGGATGAGATGGCCCGCCATCTGCGCGAGT
ACCAGGACCTGCTCAAAGTGAAGATGGCCCTGGATGTGGAGATTGCCACCTACCGGAAGCTGCTGGAGGGGAGAGGAGAGCGGTGAGGGGGCCAGGCAGGA
GCCAGTGGGAGGTGGGGGTGCTGGGTGGTCCATTTCTGTCCCAGGAGGTGAGATTACTGATTACCTCAACAAGACCTGGAAACAATTTTTTTTTT
TTTTTGAGATGGAGTTTCCTCTTGTCCCGAGTCTGGAGTGCAATGGCACCATCTTGGCTCACTGCAACCTCCGCTCCTGGGTTCAGCAATTTCTCCT
```

דוגמא: איי CpG

- נרצה למדל את רצף הדנ"א כמהלך מרקובי



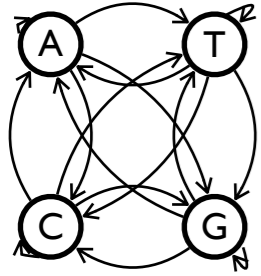
	A	C	G	T
A				
C				
G				
T				

- וכך נוכל לחשב את הניראות של הרצף

chr2:220282501-220286500
GAGGCTCAGGGCTAGCTCGCCATAGACATACATGGCAGGCAGGCTTTGGCCAGGATCCC

- נשערך מטריצה אחת עבור דנ"א בתוך איי CpG, ומטריצה נוספת עבור דנ"א מחוץ לאיים.

דוגמא: איי CpG



- נרצה למדל את רצף הדנ"א כמהלך מרקובי

הסתברויות מעבר
בתוך איי CpG

b

τ^+	A	C	G	T
A	18%	27%	43%	12%
C	17%	37%	27%	19%
G	16%	34%	38%	12%
T	8%	36%	38%	18%

a

$P_+(a \rightarrow b)$

הסתברויות מעבר
מחוץ לאיי CpG

b

τ	A	C	G	T
A	30%	20%	29%	21%
C	32%	30%	8%	30%
G	25%	25%	30%	20%
T	18%	24%	29%	29%

a

$P_-(a \rightarrow b)$

- והנידאות עבור רצף כלשהו

$$P(\vec{X} | \tau^+) = P(X_1) \cdot \prod_{i=2}^n P_{\tau^+}(X_i | X_{i-1}) = P(X_1) \cdot \prod_{i=2}^n [P_{\tau^+}]_{X_{i-1}, X_i}$$

דוגמא: CpG איי

• וכמו שעשינו בעבר, לוג יחס הנראות מתפרק לסכום

$$S(\vec{X}) = \log \frac{P(\vec{X} | \tau^+)}{P(\vec{X} | \tau^-)} = \sum_{i=1}^n \log \frac{[P_{\tau^+}]_{X_{i-1}, X_i}}{[P_{\tau^-}]_{X_{i-1}, X_i}} = \sum_{i=1}^n \sigma_{X_{i-1}, X_i}$$

σ	A	C	G	T
A	-0.74	0.42	0.58	-0.8
C	-0.91	0.3	1.8	-0.68
G	-0.62	0.46	0.33	-0.73
T	-0.12	0.57	0.39	-0.68

מטריצת סקור
[לזוגות עוקבים]

τ^+	A	C	G	T
A	18%	27%	43%	12%
C	17%	37%	27%	19%
G	16%	34%	38%	12%
T	8%	36%	38%	18%

= log (

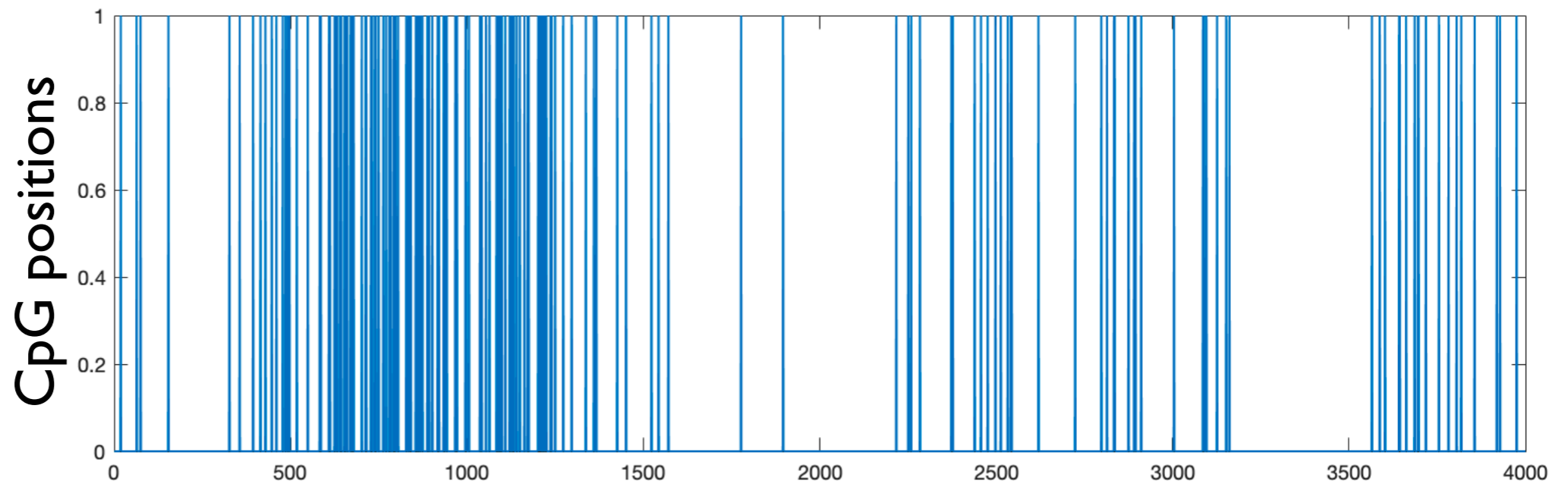
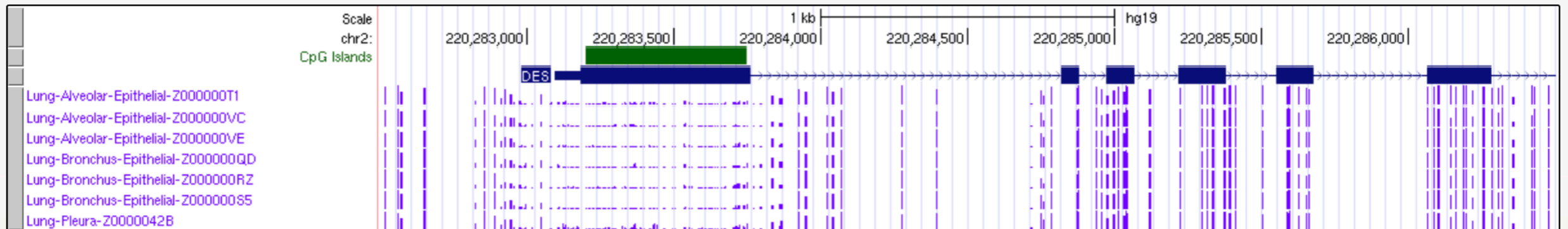
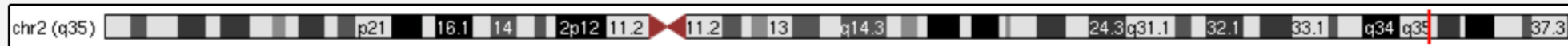
τ^-	A	C	G	T
A	30%	20%	29%	21%
C	32%	30%	8%	30%
G	25%	25%	30%	20%
T	18%	24%	29%	29%

CpG Islands

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr2:220,282,501-220,286,500 4,000 bp. enter position, gene symbol, HGVS or search terms go

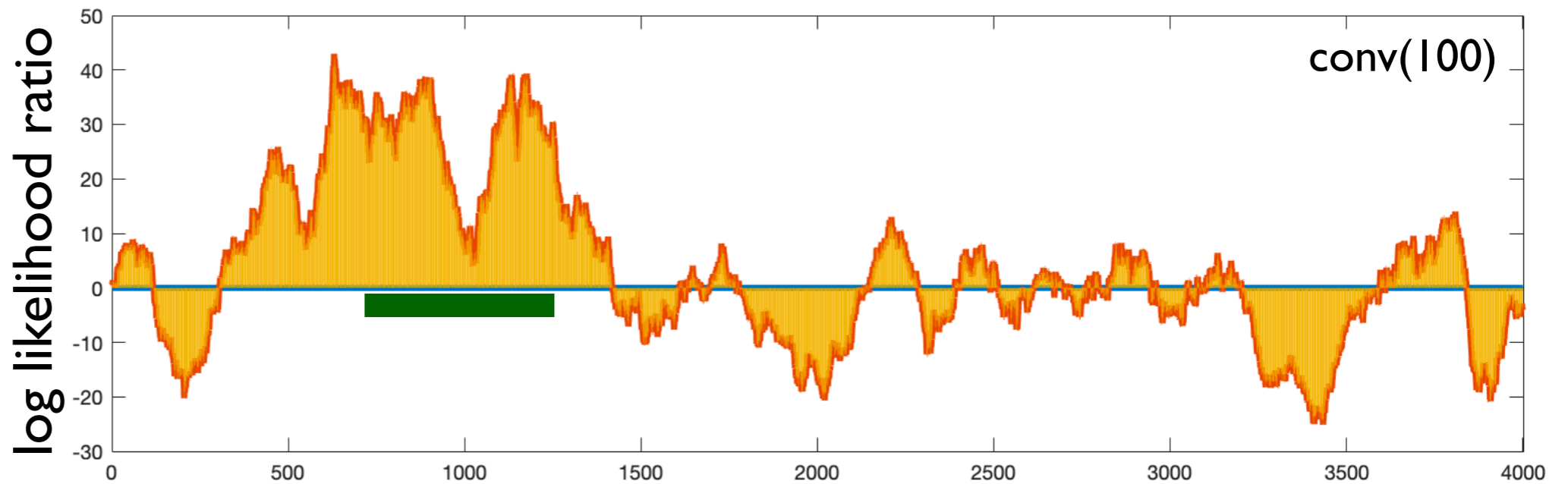
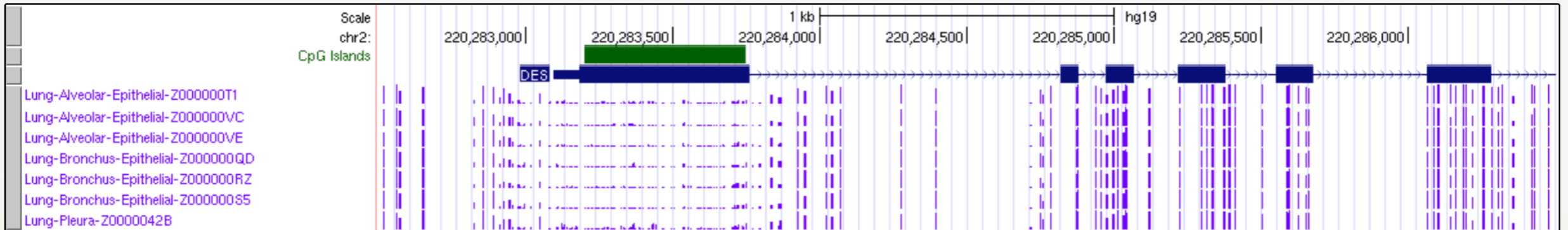
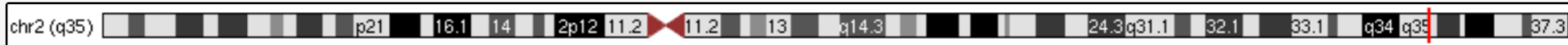


CpG Islands

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

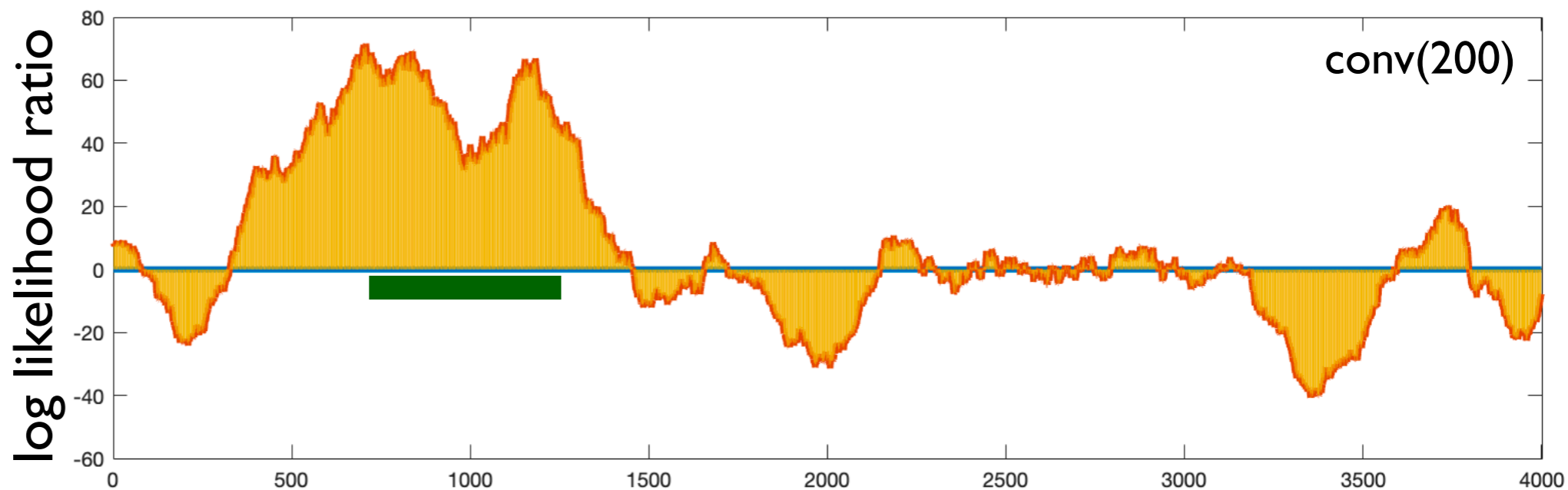
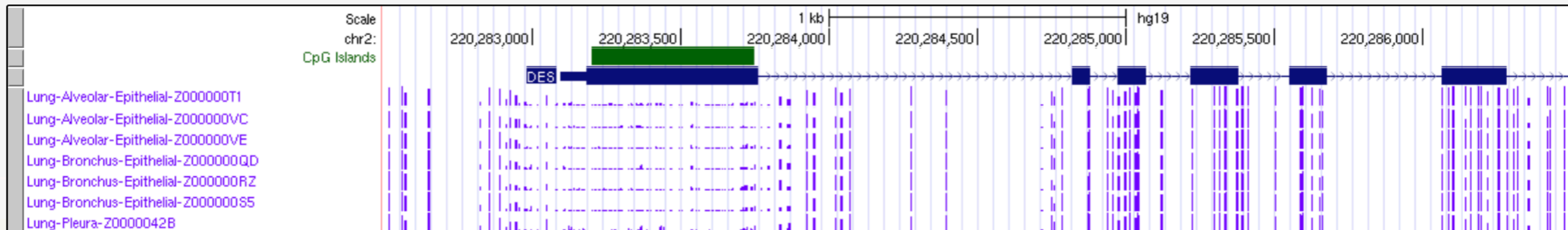
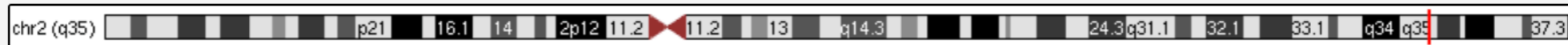
chr2:220,282,501-220,286,500 4,000 bp. enter position, gene symbol, HGVS or search terms go



CpG Islands

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

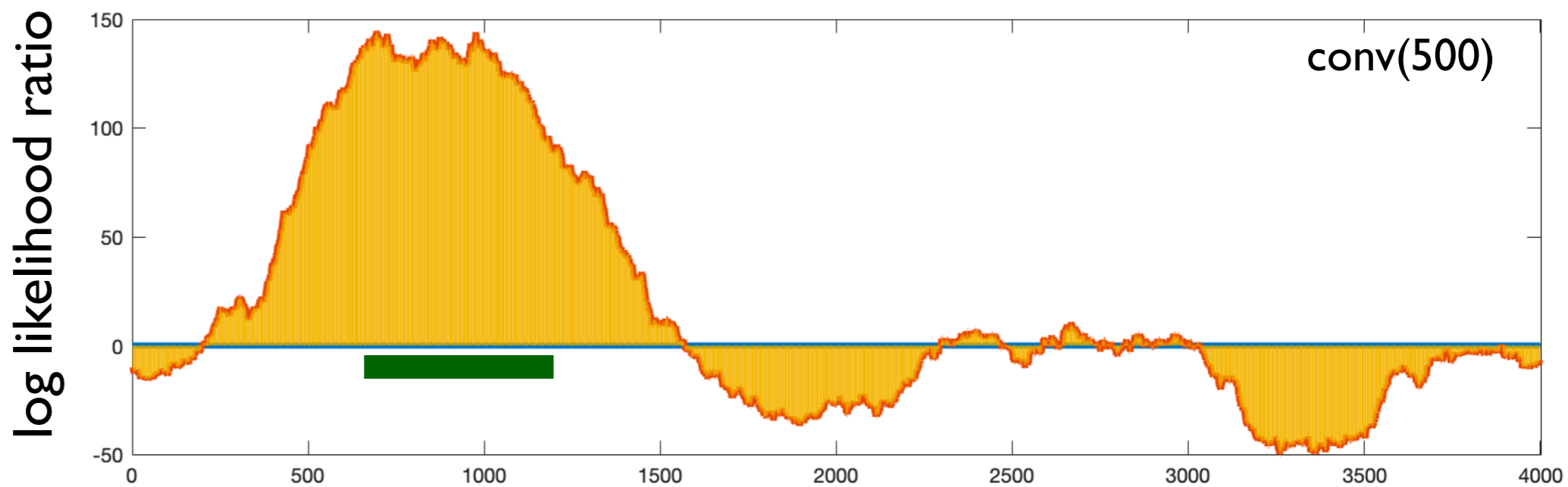
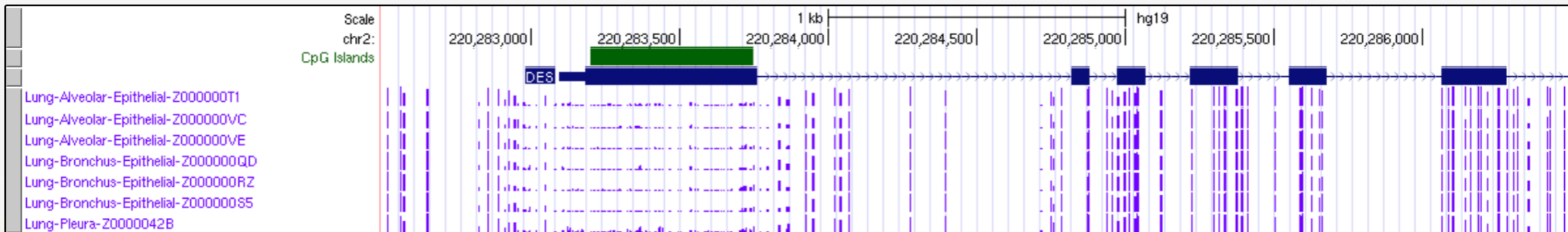
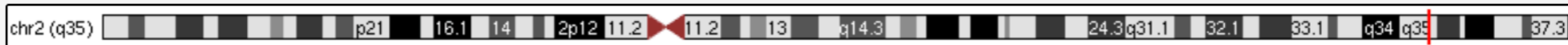
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr2:220,282,501-220,286,500 4,000 bp. enter position, gene symbol, HGVS or search terms go



CpG Islands Analysis

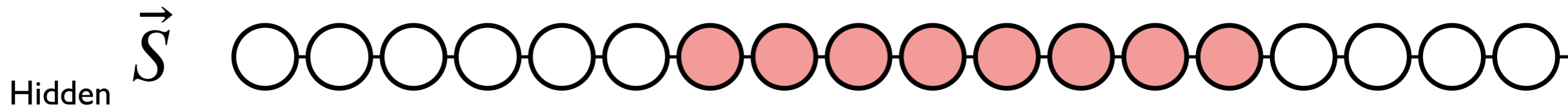
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr2:220,282,501-220,286,500 4,000 bp. enter position, gene symbol, HGVS or search terms go

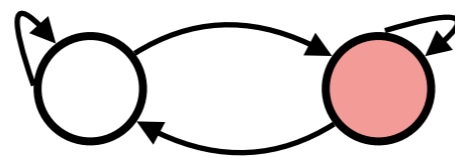


שרשראות מרקוב חבויות

- משתני התהליך המרקוביים הם חבויים



$$S_i \in \{\text{Fair}, \text{Loaded}\}$$

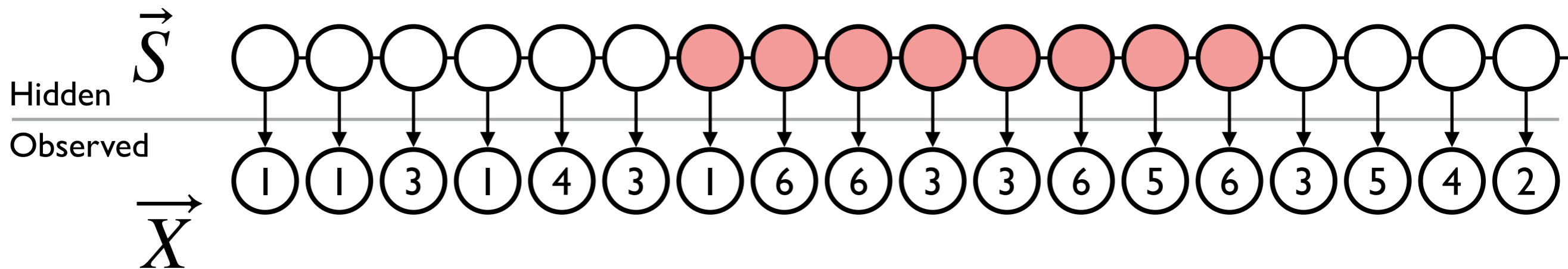


τ

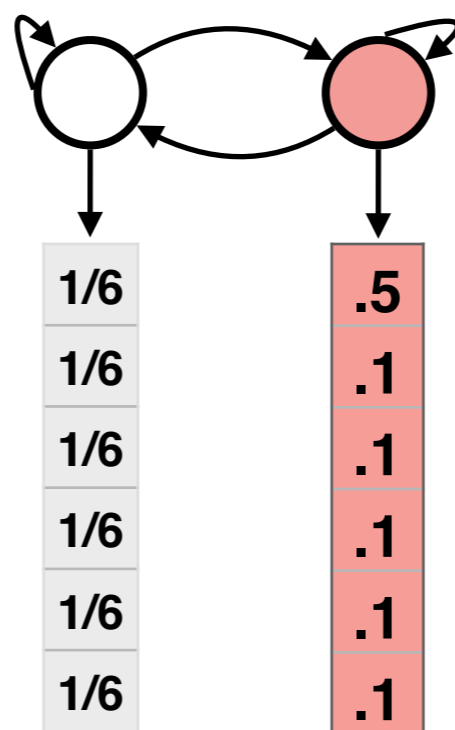
τ	F	L
F	99%	1%
L	5%	95%

שרשראות מרקוב חבויות

- משתני התהליך המרקוביים הם חבויים
- פולטים (בהתאם למצבם) משתנים אחרים (נצפים)



$$S_i \in \{\text{Fair}, \text{Loaded}\}$$



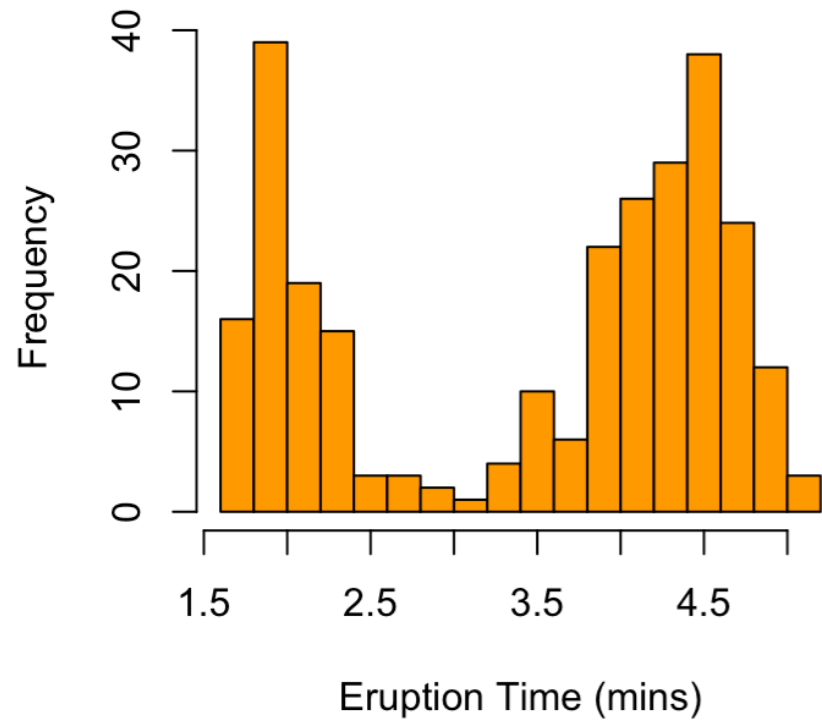
τ

τ	F	L
F	99%	1%
L	5%	95%

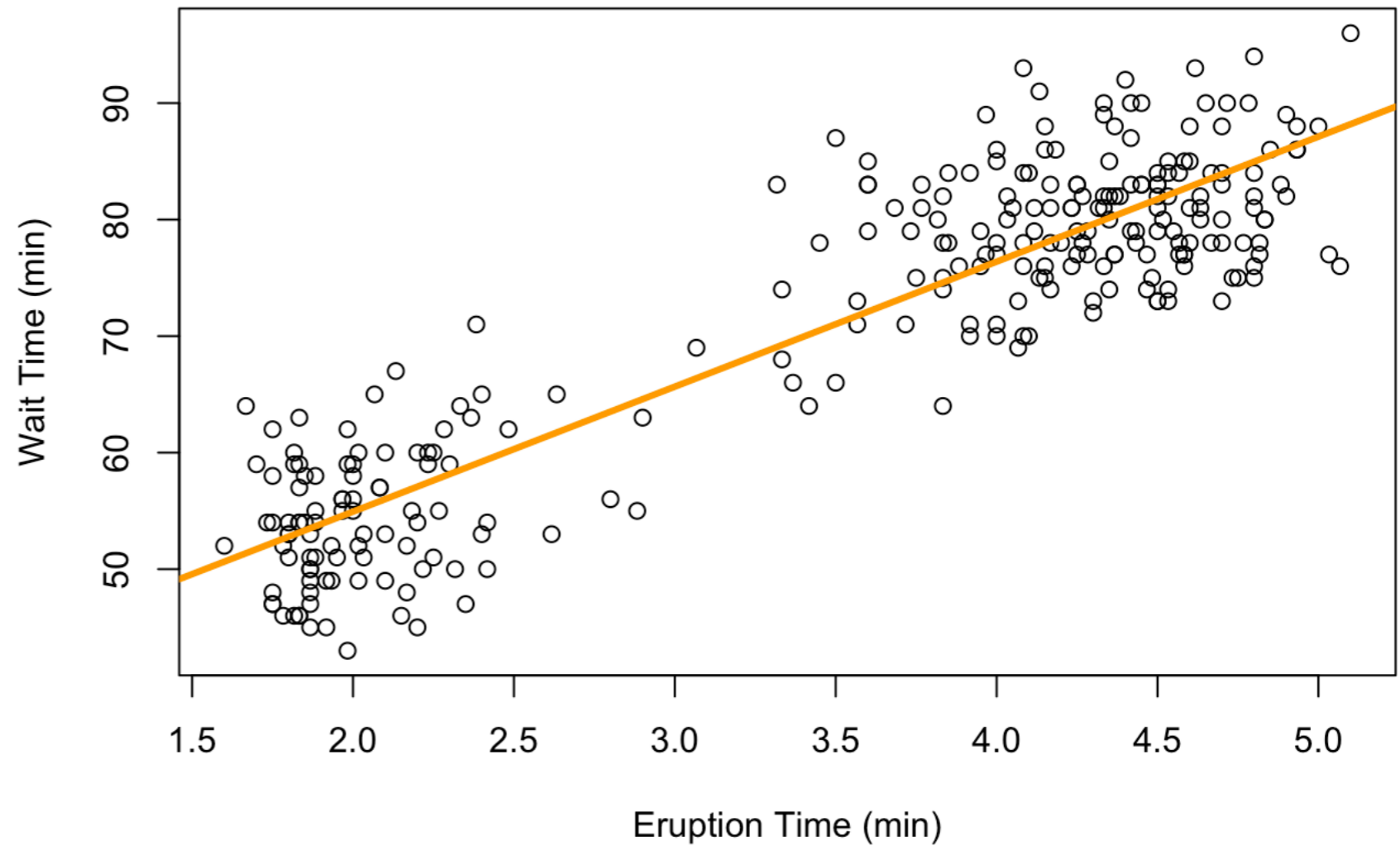
$$X_i \in \{1, 2, 3, 4, 5, 6\}$$

הגייזר אולד פוייט'פול

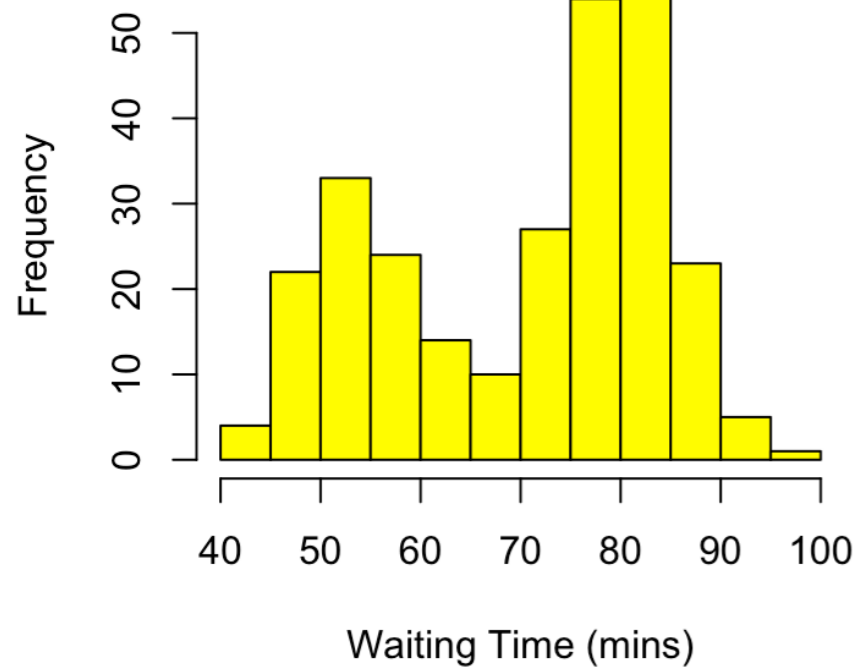
Frequency of the Eruption Duration



Dispersion of Old Faithful Eruptions



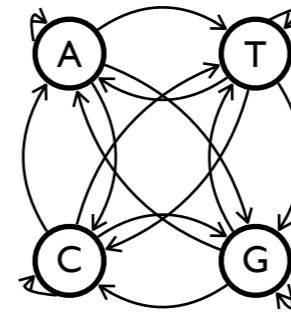
Frequency of Time Waiting



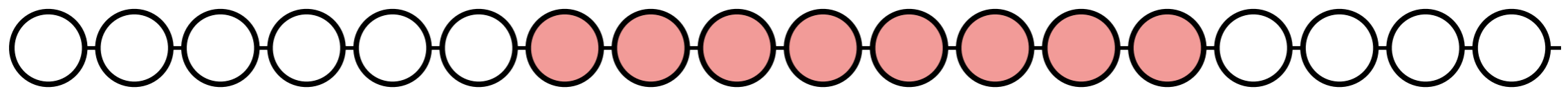
דיון - מודל מרקובי מסדר גבוה

• מודל מסדר ראשון

$$P(\vec{X} | \tau^+) = P(X_1) \cdot \prod_{i=2}^n P_{\tau^+}(X_i | X_{i-1})$$

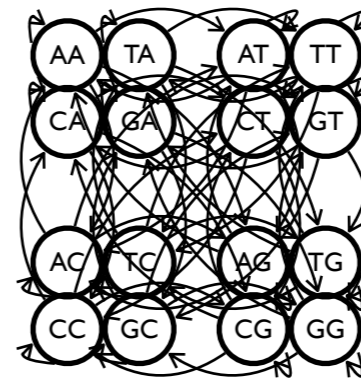


	A	C	G	T
A				
C				
G				
T				

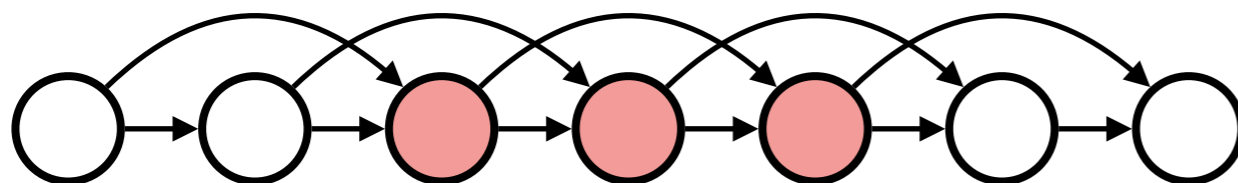


• מודל מסדר שני

$$P(\vec{X} | \tau^+) = P(X_1) \cdot P(X_2 | X_1) \cdot \prod_{i=3}^n P(X_i | X_{i-2}, X_{i-1})$$



	A	C	G	T
AA				
CA				
GA				
TA				
AC				
CC				
GC				
TC				
...
TT				



דיון - מודל מרקובי מסדר גבוה

• מודל מרקובי מסדר משתנה

Machine Learning, 25, 117–149 (1996)

The Power of Amnesia: Learning Probabilistic Automata with Variable Memory Length

DANA RON

Laboratory for Computer Science, MIT, Cambridge, MA 02139

danar@theory.lcs.mit.edu

YORAM SINGER

AT&T Labs, 600 Mountain Avenue, Murray Hill, NJ 07974

singer@research.att.com

NAFTALI TISHBY

Institute of Computer Science, Hebrew University, Jerusalem 91904, Israel

tishby@cs.huji.ac.il

Editor: Thomas Hancock

Abstract. We propose and analyze a distribution learning algorithm for **variable memory length Markov processes**. These processes can be described by a subclass of probabilistic finite automata which we name **Probabilistic Suffix Automata (PSA)**. Though hardness results are known for learning distributions generated by general probabilistic automata, we prove that the algorithm we present can efficiently learn distributions generated by PSAs. In particular, we show that for any target PSA, the KL-divergence between the distribution generated by the target and the distribution generated by the hypothesis the learning algorithm outputs, can be made small with high confidence in polynomial time and sample complexity. The learning algorithm is motivated by applications in human-machine interaction. Here we present two applications of the algorithm. In the first one we apply the algorithm in order to construct a model of the English language, and use this model to correct corrupted text. In the second application we construct a simple stochastic model for *E.coli* DNA.