האוניברסיטה
העברית
בירושלים
THE HEBREW
UNIVERSITY
OF JERUSALEM

אלגוריתמים בביו' חישובית

76558

מודלים מרקוביים
ושרשראות מרקוב חבויות

תומי קפלן

16/1/2024

מודלים מרקוביים חבויים

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

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

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

תכונת המרקוביות

- ניזכר בכלל השרשרת:

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

- או

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

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

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

תכונת המרקוביות

- בלומר, תת-הסדרה $X_{i+1},\ldots,X_j$ אינה תלויה בערכי $X_1, \ldots, X_{i-1}$ בהנתן X_i

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

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

- ותחת הנחת המרקוביות, נוכל לשכתב

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

תכונת המרקוביות

- בלומר, תת-הסדרה $X_{i+1},\ldots,X_j$ אינה תלויה בערכי $X_1, \ldots, X_{i-1}$ בהנתן X_i

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

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

- ותחת הנחת המרקוביות, נוכל לשכתב

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

תהליך מרקובי

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

- נמדל את התהליך בעזרת:
  - וקטור ההתפלגות ההתחלתית $X_1 \sim \overrightarrow{P_0}$
  - מטריצת מעברים $\left[P_t\right]_{i,j} = P(X_t = S_j \,|\, X_{t-1} = S_i)$

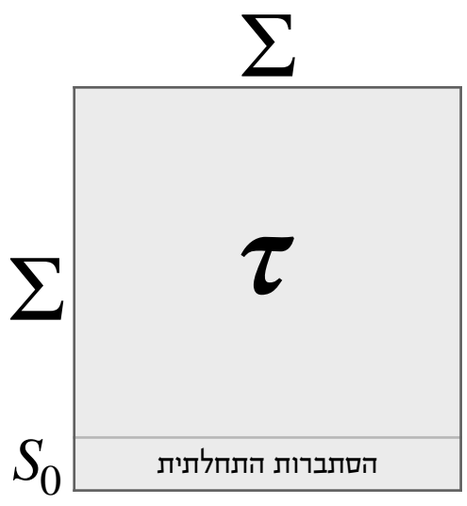
- בד"כ נניח שהתהליך הומוגני בזמן

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

מטריצת הסתברויות מעבר

- נגדיר: $\tau_{a,b} \triangleq P(X_t = b \,|\, X_{t-1} = a) = P(a \to b)$
- וכמובן $\forall_{a,b} \;\; \tau_{a,b} \geq 0 \qquad \forall_a \;\; \sum_b \tau_{a,b} = 1$
- לפעמים שימושי להוסיף מצב התחלה [פיקטיבי] ולהגדיר את וקטור ההתפלגות ההתחלתית בשורה נוספת ב- τ

$S_0 \notin \Sigma$

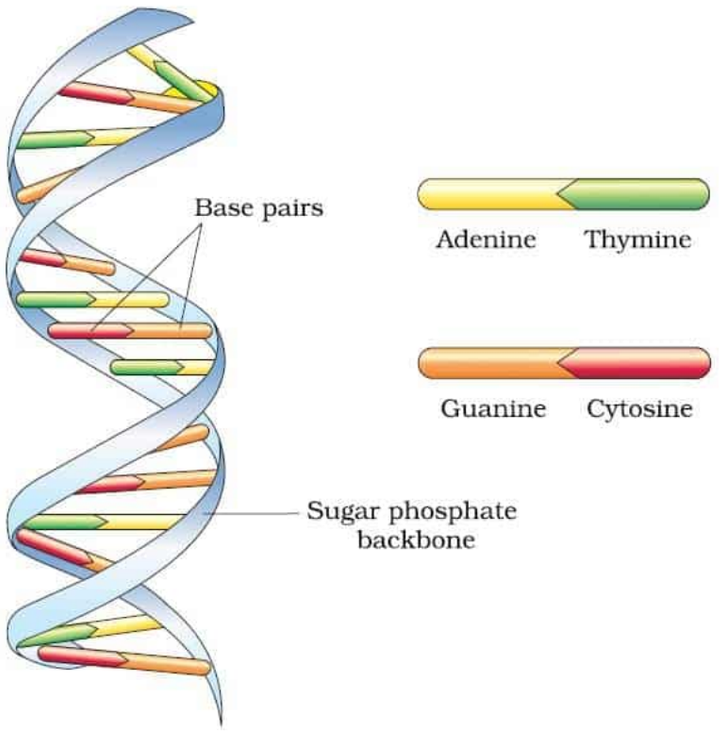


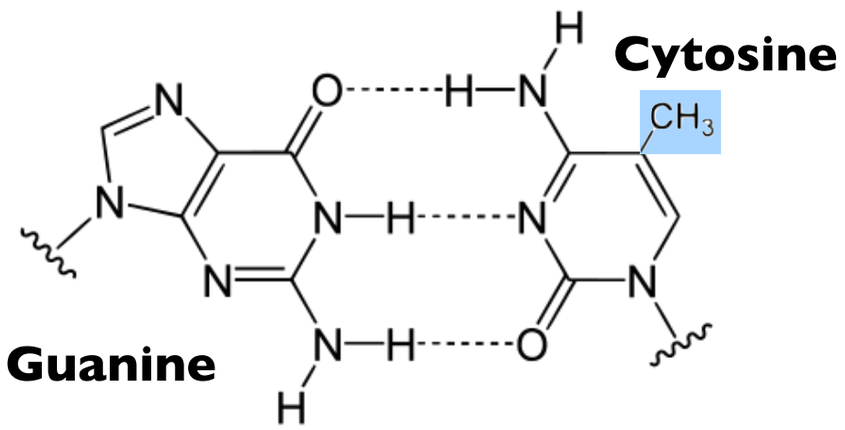
דוגמא: איי CpG

- בגנום, באשר ציטוזין מופיע לפני גואנין,
ב-80% מהמקרים הציטוזין יעבור מתילציה
ל- 5-מתיל-ציטוזין

5′— C-G —3′
3′— G-C —5′

- המתילציה תישמר לאורך כל חיי התא, ותגדיר
את ה"זהות התאית" שלו ואיזה גנים הוא
יבטא או ישתיק

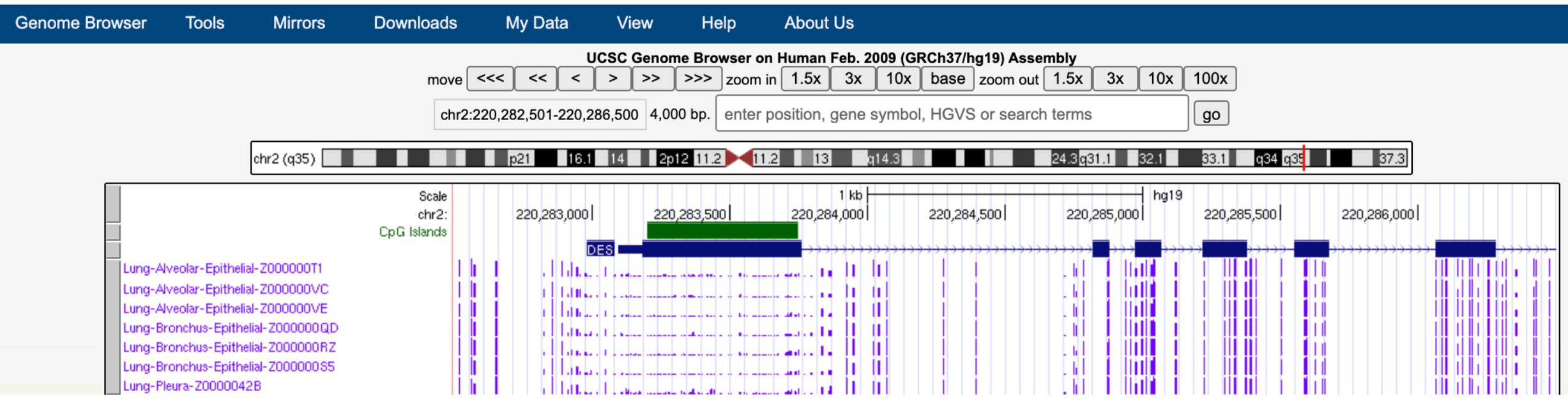




דוגמא: איי CpG

- 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%]



דוגמא: איי CpG

```
>chr2:220282501-220286500
GAGGCTCAGGGCTAGCTCGCCCATAGACATACATGGCAGGCAGGCTTTGGCCAGGATCCCTCCGCCTGCCAGGCGTCTCCCTGCCCTCCCTTCCTGCCTA
GAGACCCCCACCCTCAAGCCTGGCTGGTCTTTGCCTGAGACCCAAACCTCTTCGACTTCAAGAGAATATTTAGGAACAAGGTGGTTTAGGGCCTTTCCTG
GGAACAGGCCTTGACCCTTTAAGAAATGACCCAAAGTCTCTCCTTGACCAAAAAGGGGACCCTCAAACTAAAGGGAAGCCTCTCTTCTGCTGTCTCCCCT
GACCCCACTCCCCCCCACCCCAGGACGAGGAGATAACCAGGGCTGAAAGAGGCCCGCCTGGGGGCTGCAGACATGCTTGCTGCCTGCCCTGGCGAAGGAT
TGGCAGGCTTGCCCGTCACAGGACCCCCGCTGGCTGACTCAGGGGCGCAGGCCTCTTGCGGGGGAGCTGGCCTCCCCGCCCCCACGGCCACGGGCCGCCC
TTTCCTGGCAGGACAGCGGGATCTTGCAGCTGTCAGGGGAGGGGAGGCGGGGGCTGATGTCAGGAGGGATACAAATAGTGCCGACGGCTGGGGGCCCTGT
CTCCCCTCGCCGCATCCACTCTCCGGCCGGCCGCCTGCCCGCCGCCTCCTCCGTGCGCCCGCCAGCCTCGCCCGCGCCGTCACCATGAGCCAGGCCTACT
CGTCCAGCCAGCGCGTGTCCTCCTACCGCCGCACCTTCGGCGGGGCCCCGGGCTTCCCACTCGGCTCCCCGCTGAGTTCGCCCGTGTTCCCGCGGGCGGG
TTTCGGCTCTAAGGGCTCCTCCAGCTCGGTGACGTCCCGCGTGTACCAGGTGTCGCGCACGTCGGGCGGGGCCGGGGGCCTGGGGTCGCTGCGGGCCAGC
CGGCTGGGGACCACCCGCACGCCCTCCTCCTACGGCGCAGGCGAGCTGCTGGACTTCTCACTGGCCGACGCGGTGAACCAGGAGTTTCTGACCACGCGCA
CCAACGAGAAGGTGGAGCTGCAGGAGCTCAATGACCGCTTCGCCAACTACATCGAGAAGGTGCGCTTCCTGGAGCAGCAGAACGCGGCGCTCGCCGCCGA
AGTGAACCGGCTCAAGGGCCGCGAGCCGACGCGAGTGGCCGAGCTCTACGAGGAGGAGCTGCGGGAGCTGCGGCGCCAGGTGGAGGTGCTCACTAACCAG
CGCGCGCGCGTCGACGTCGAGCGCGACAACCTGCTCGACGACCTGCAGCGGCTCAAGGCCAAGTGAGGGCCCGGCACCCCAGACTCCTCTTTCTGCGGGC
AGGGCACAGGAGGCTAGGCCTGGGGTCTGGGGTCCCGCTGTCAGCACCTGCCTTCTCCCGGGGCCCGGGACCCTCTCCTGCCCCATGTGGAGAAAGGGTC
CTCCACCTGTGTGTTTCAAGGGCCGTGACCTCCAGGTCTCTCCCCCTGCGATCCCATCTTGCACAGGAGTTTTCTTGGGGACATAGATCAGGGGGTGGA
TATGGGAGAATTTAGGGGACCCGGTGCCCTGTGGACAGCCCCGTTAAAAAGCATTTTAAGATGCTGGGGCGATATTTATGGGGTCAGGTAGTTGATGGGC
AGAGGAAGGGCTGCAGGAGGCCCAGAGGGCAGTGTAGCCAGAGGGAGAAGGGAGGCTGATAGGAGACAGGGAAAGCAGGGCAAGGGCCCAGAGTCCAAGC
AACAGCTCTCAGCTCAGCTGTGATGAGGCCCTGGGGGAGGTGGGGGGAGGGGGGAGCTTGGCCCTGGGGCCTTGCCGAGACTGTGTCTTTTTACAAGGTG
AATGGACAGGCTGGAGAAAAAGGGAGTAGGTGGGGGTCACAGCTCTCAGAGAGCTTGGGAGGACCTGACTGTAGACTTCACCAGGCTCCAAGAACGAAAA
GGGCAGCAAGTGTAGCATATTTGTTGGTCCCACTTCTGACAGGCCAAGTGAGCACAGTCACCCTCCTGCCACCAAAGTCATAAATATTAATTGAGCAGCT
ATATTGGCCAGGCTGGAGCTGGGAACCAGAAACACAGAGGTGGATAAAATAGACACAGTTTCTAACCCCAGGGAGGTCACACAGTCTGGTGGGGACATAG
ACTTCAAGGGTGTGGCTCCTGGGCAGAGATTGGGCCACTTCCTGTGCCCTCCCTGGGTGGGTGGGGCCTCTCCACTCCCTGTCTCTCCTGCCTCTACCCA
GCAGCCAGGCCCTCCCGCTCTGTCCTGGACCCACCCCCTGGTCAGCCCCCGGCCAGTCGTTTCCACTGCCAGCTTTATCACCCGCAACTGTCTGTCTTTC
TGTCTGTCCCACCCAGGCTGCAGGAGGAGATTCAGTTGAAGGAAGAAGCAGAGAACAATTTGGCTGCCTTCCGAGCGGTGAGTGCCCTTCTTTTCCCCTT
GCATGGCCTCTGGCCTTGCTCTGCCCCACCTGGGTGGCGGTGACCATGTCCTTCTCGCTTGGCCTCTCCCAGGACGTGGATGCAGCTACTCTAGCTCGCA
TTGACCTGGAGCGCAGAATTGAATCTCTCAACGAGGAGATCGCGTTCCTTAAGAAAGTGCATGAAGAGGTATACCTTGGCCCCTCTTCCTGGGGTCACTG
GGCCATGGGGAAAGCAGCCGGAAAGTGGGGTTGGGGTGAGGCTCTGGCTGGGAATAGGGGTGTGAGGGTGCTGTGTGGGCCCTGAGAGGGGACTGAAGCC
CAGTCATGCCCTACAGGAGATCCGTGAGTTGCAGGCTCAGCTTCAGGAACAGCAGGTCCAGGTGGAGATGGACATGTCTAAGCCAGACCTCACTGCCGCC
CTCAGGGACATCCGGGCTCAGTATGAGACCATCGCGGCTAAGAACATTTCTGAAGCTGAGGAGTGGTACAAGTCGAAGGTGGGTGGCCTCGCCCGGGGAC
TGGCATCTCCGTCCCCCTGAATCCCAGCTTGGATGTGCTGCCTGTGGTACCATCCATGGGAGGAGAGCCCAGAGGCTTCATGCTCCCTTGCTCATCCCTA
CCCGTGCCCTGCATCCTTCTCATTTTTGGGCCCCTTTCTCTGCCCTTAGGTGTCAGACCTGACCCAGGCAGCCAACAAGAACAACGACGCCCTGCGCCAG
GCCAAGCAGGAGATGATGGAATACCGACACCAGATCCAGTCCTACACCTGCGAGATTGACGCCCTGAAGGGCACTGTGAGTCCCTGCCCACCTGGCCAGG
CCCTGCCCCTTCCTGTCTGCAGTTCACACCCTCACTTTGTGACCTTGGGCCCATCATAGATCCTCTCTGGGCCTTCATCTACTTAAATCTACAATAGGGG
TAAAACCAGACAAGTGGATTCCAGTTGGATGCTAAGGAATCAGGGGTTCCTGGGCATCTACCTATGTGGGGACTGTGAGGCTGAATGCAATGTTCCTTTG
TATCTATTTTATTCTGAGTGTTCACATATAGACTTAATTTGAGTTCAGGGTTCAACATGGCCTGGACCTGACCATCTGGAGTTGCCTGCCAGCCCCAAAG
CTTTCTTTGGGCTGCTAGTGTCCTCTTCCCTTCCTTGACCTGGGTTCCCCCTCTCCTGCAGAACGATTCCCTGATGAGGCAGATGCGGGAATTGGAGGAC
CGATTTGCCAGTGAGGCCAGTGGCTACCAGGACAACATTGCGCGCCTGGAGGAGGAAATCCGGCACCTCAAGGATGAGATGGCCCGCCATCTGCGCGAGT
ACCAGGACCTGCTCAACGTGAAGATGGCCCTGGATGTGGAGATTGCCACCTACCGGAAGCTGCTGGAGGGAGAGGAGAGCCGGTGAGGGGCCAGGCAGGA
GCCCGAGTGGGAGGTGCGGGGTGCTGGGTGGTCCATTTCTGTCCCCAGGAGGCTCGAGATTACTGATTACCTCAACAAGACCTGGAAACAATTTTTTTTT
TTTTTGAGATGGAGTTTCGCTCTTGTCGCCCAGTCTGGAGTGCAATGGCACCATCTTGGCTCACTGCAACCTCCGCCTCCTGGGTTCAAGCAATTCTCCT
```

GC

דוגמא: איי CpG

>chr2:220282501-220286500

GAGGCTCAGGGCTAGCTCGCCCATAGACATACATGGCAGGCAGGCTTTGGCCAGGATCCCTCCGCCTGCCAGGCGTCTCCCTGCCCTCCCTTCCTGCCTA
GAGACCCCCACCCTCAAGCCTGGCTGGTCTTTGCCTGAGACCCAAACCTCTTCGACTTCAAGAGAATATTTAGGAACAAGGTGGTTTAGGGCCTTTCCTG
GGAACAGGCCTTGACCCTTTAAGAAATGACCCAAAGTCTCTCCTTGACCAAAAAGGGGACCCTCAAACTAAAGGGAAGCCTCTCTTCTGCTGTCTCCCCT
GACCCCACTCCCCCCCACCCCAGGACGAGGAGATAACCAGGGCTGAAAGAGGCCCGCCTGGGGGCTGCAGACATGCTTGCTGCCTGCCCTGGCGAAGGAT
TGGCAGGCTTGCCCGTCACAGGACCCCCGCTGGCTGACTCAGGGGCGCAGGCCTCTTGCGGGGGAGCTGGCCTCCCCGCCCCCACGGCCACGGGCCGCCC
TTTCCTGGCAGGACAGCGGGATCTTGCAGCTGTCAGGGGAGGGGAGGCGGGGGCTGATGTCAGGAGGGATACAAATAGTGCCGACGGCTGGGGGCCCTGT
CTCCCCTCGCCGCATCCACTCTCCGGCCGGCCGCCTGCCCGCCGCTCCTCCGTGCGCCCGCCAGCCTCGCCCGCGCCGTCACCATGAGCCAGGCCTACT
CGTCCAGCCAGCGCGTGTCCTCCTACCGCCGCACCTTCGGCGGGGCCCCGGGCTTCCCACTCGGCTCCCCGCTGAGTTCGCCCGTGTTCCCGCGGGCGGG
TTTCGGCTCTAAGGGCTCCTCCAGCTCGGTGACGTCCCGCGTGTACCAGGTGTCGCGCACGTCGGGCGGGGCCGGGGGCCTGGGGTCGCTGCGGGCCAGC
CGGCTGGGGACCACCCGCACGCCCTCCTCCTACGGCGCAGGCGAGCTGCTGGACTTCTCACTGGCCGACGCGGTGAACCAGGAGTTTCTGACCACGCGCA
CCAACGAGAAGGTGGAGCTGCAGGAGCTCAATGACCGCTTCGCCAACTACATCGAGAAGGTGCGCTTCCTGGAGCAGCAGAACGCGGCGCTCGCCGCCGA
AGTGAACCGGCTCAAGGGCCGCGAGCCGACGCGAGTGGCCGAGCTCTACGAGGAGGAGCTGCGGGAGCTGCGGCGCCAGGTGGAGGTGCTCACTAACCAG
CGCGCGCGCGTCGACGTCGAGCGCGACAACCTGCTCGACGACCTGCAGCGGCTCAAGGCCAAGTGAGGGCCCGGCACCCCAGACTCCTCTTTCTGCGGGC
AGGGCACAGGAGGCTAGGCCTGGGGTCTGGGGTCCCGCTGTCAGCACCTGCCTTCTCCCGGGGCCCGGGACCCTCTCCTGCCCCATGTGGAGAAAGGGTC
CTCCACCTGTGTGTTTCAAGGGCCGTGACCTCCAGGTCTCTCCCCCTGCGATCCCATCTTGCACAGGAGTTTTCTTGGGGACATAGATCAGGGGGTGGA
TATGGGAGAATTTAGGGGACCCGGTGCCCTGTGGACAGCCCCGTTAAAAAGCATTTTAAGATGCTGGGGCGATATTTATGGGGTCAGGTAGTTGATGGGC
AGAGGAAGGGCTGCAGGAGGCCCAGAGGGCAGTGTAGCCAGAGGGAGAAGGGAGGCTGATAGGAGACAGGGAAAGCAGGGCAAGGGCCCAGAGTCCAAGC
AACAGCTCTCAGCTCAGCTGTGATGAGGCCCTGGGGGAGGTGGGGGGAGGGGGGAGCTTGGCCCTGGGGCCTTGCCGAGACTGTGTCTTTTTACAAGGTG
AATGGACAGGCTGGAGAAAAAGGGAGTAGGTGGGGGTCACAGCTCTCAGAGAGCTTGGGAGGACCTGACTGTAGACTTCACCAGGCTCCAAGAACGAAAA
GGGCAGCAAGTGTAGCATATTTGTTGGTCCCACTTCTGACAGGCCAAGTGAGCACAGTCACCCTCCTGCCACCAAAGTCATAAATATTAATTGAGCAGCT
ATATTGGCCAGGCTGGAGCTGGGAACCAGAAACACAGAGGTGGATAAAATAGACACAGTTTCTAACCCCAGGGAGGTCACACAGTCTGGTGGGGACATAG
ACTTCAAGGGTGTGGCTCCTGGGCAGAGATTGGGCCACTTCCTGTGCCCTCCCTGGGTGGGTGGGGCCTCTCCACTCCCTGTCTCTCCTGCCTCTACCCA
GCAGCCAGGCCCTCCCGCTCTGTCCTGGACCCACCCCCTGGTCAGCCCCCGGCCAGTCGTTTCCACTGCCAGCTTTATCACCCGCAACTGTCTGTCTTTC
TGTCTGTCCCACCCAGGCTGCAGGAGGAGATTCAGTTGAAGGAAGAAGCAGAGAACAATTTGGCTGCCTTCCGAGCGGTGAGTGCCCTTCTTTTCCCCTT
GCATGGCCTCTGGCCTTGCTCTGCCCCACCTGGGTGGCGGTGACCATGTCCTTCTCGCTTGGCCTCTCCCAGGACGTGGATGCAGCTACTCTAGCTCGCA
TTGACCTGGAGCGCAGAATTGAATCTCTCAACGAGGAGATCGCGTTCCTTAAGAAAGTGCATGAAGAGGTATACCTTGGCCCCTCTTCCTGGGGTCACTG
GGCCATGGGGAAAGCAGCCGGAAAGTGGGGTTGGGGTGAGGCTCTGGCTGGGAATAGGGGTGTGAGGGTGCTGTGTGGGCCCTGAGAGGGGACTGAAGCC
CAGTCATGCCCTACAGGAGATCCGTGAGTTGCAGGCTCAGCTTCAGGAACAGCAGGTCCAGGTGGAGATGGACATGTCTAAGCCAGACCTCACTGCCGCC
CTCAGGGACATCCGGGCTCAGTATGAGACCATCGCGGCTAAGAACATTTCTGAAGCTGAGGAGTGGTACAAGTCGAAGGTGGGTGGCCTCGCCCGGGGAC
TGGCATCTCCGTCCCCCTGAATCCCAGCTTGGATGTGCTGCCTGTGGTACCATCCATGGGAGGAGAGCCCAGAGGCTTCATGCTCCCTTGCTCATCCCTA
CCCGTGCCCTGCATCCTTCTCATTTTTGGGCCCCTTTCTCTGCCCTTAGGTGTCAGACCTGACCCAGGCAGCCAACAAGAACAACGACGCCCTGCGCCAG
GCCAAGCAGGAGATGATGGAATACCGACACCAGATCCAGTCCTACACCTGCGAGATTGACGCCCTGAAGGGCACTGTGAGTCCCTGCCCACCTGGCCAGG
CCCTGCCCCTTCCTGTCTGCAGTTCACACCCTCACTTTGTGACCTTGGGCCCATCATAGATCCTCTCTGGGCCTTCATCTACTTAAATCTACAATAGGGG
TAAAACCAGACAAGTGGATTCCAGTTGGATGCTAAGGAATCAGGGGTTCCTGGGCATCTACCTATGTGGGGACTGTGAGGCTGAATGCAATGTTCCTTTG
TATCTATTTTATTCTGAGTGTTCACATATAGACTTAATTTGAGTTCAGGGTTCAACATGGCCTGGACCTGACCATCTGGAGTTGCCTGCCAGCCCCAAAG
CTTTCTTTGGGCTGCTAGTGTCCTCTTCCCTTCCTTGACCTGGGTTCCCCCTCTCCTGCAGAACGATTCCCTGATGAGGCAGATGCGGGAATTGGAGGAC
CGATTTGCCAGTGAGGCCAGTGGCTACCAGGACAACATTGCGCGCCTGGAGGAGGAAATCCGGCACCTCAAGGATGAGATGGCCGCCATCTGCGCGAGT
ACCAGGACCTGCTCAACGTGAAGATGGCCCTGGATGTGGAGATTGCCACCTACCGGAAGCTGCTGGAGGGAGAGGAGAGCCGGTGAGGGGCCAGGCAGGA
GCCGAGTGGGAGGTGCGGGGTGCTGGGTGGTCCATTTCTGTCCCCAGGAGGCTCGAGATTACTGATTACCTCAACAAGACCTGGAAACAATTTTTTTTT
TTTTTGAGATGGAGTTTCGCTCTTGTCGCCCAGTCTGGAGTGCAATGGCACCATCTTGGCTCACTGCAACCTCCGCCTCCTGGGTTCAAGCAATTCTCCT

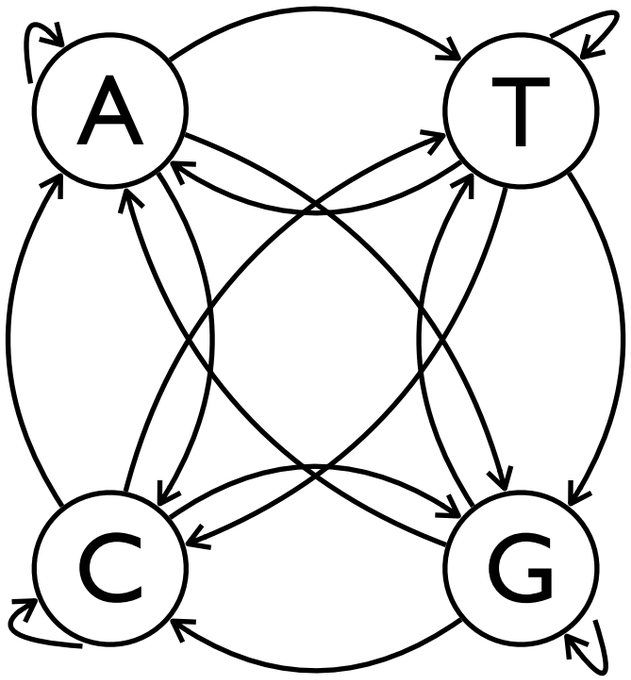
CG

CpG דוגמא: איי

>chr2:220282501-220286500
GAGGCTCAGGGCTAGCTCGCCCATAGACATACATGGCAGGCAGGCTTTGGCCAGGATCCCTCCGCCTGCCAGGCGTCTCCCTGCCCTCCCTTCCTGCCTA
GAGACCCCCACCCTCAAGCCTGGCTGGTCTTTGCCTGAGACCCAAACCTCTTCGACTTCAAGAGAATATTTAGGAACAAGGTGGTTTAGGGCCTTTCCTG
GGAACAGGCCTTGACCCTTTAAGAAATGACCCAAAGTCTCTCCTTGACCAAAAAGGGGACCCTCAAACTAAAGGGAAGCCTCTCTTCTGCTGTCTCCCCT
GACCCCACTCCCCCCCACCCCAGGACGAGGAGATAACCAGGGCTGAAAGAGGCCCGCCTGGGGGCTGCAGACATGCTTGCTGCCTGCCCTGGCGAAGGAT
TGGCAGGCTTGCCCGTCACAGGACCCCCGCTGGCTGACTCAGGGGCGCAGGCCTCTTGCGGGGGAGCTGGCCTCCCCGCCCCCACGGCCACGGGCCGCCC
TTTCCTGGCAGGACAGCGGGATCTTGCAGCTGTCAGGGGAGGGGAGGCGGGGGCTGATGTCAGGAGGGATACAAATAGTGCCGACGGCTGGGGGCCCTGT
CTCCCCTCGCCGCATCCACTCTCCGGCCGGCCGCCTGCCCGCCGCCTCCTCCGTGCGCCCGCCAGCCTCGCCCGCGCCGTCACCATGAGCCAGGCCTACT
CGTCCAGCCAGCGCGTGTCCTCCTACCGCCGCACCTTCGGCGGGGCCCCGGGCTTCCCACTCGGCTCCCCGCTGAGTTCGCCCGTGTTCCCGCGGCCGGG
TTTCGGCTCTAAGGGCTCCTCCAGCTCGGTGACGTCCCGCGTGTACCAGGTGTCGCGCACGTCGGGCGGGGCCGGGGGCCTGGGGTCGCTGCGGGCCAGC
CGGCTGGGGACCACCCGCACGCCCTCCTCCTACGGCGCAGGCGAGCTGCTGGACTTCTCACTGGCCGACGCGGTGAACCAGGAGTTTCTGACCACGCGCA
CCAACGAGAAGGTGGAGCTGCAGGAGCTCAATGACCGCTTCGCCAACTACATCGAGAAGGTGCGCTTCCTGGAGCAGCAGAACGCGGCGCTCGCCGCCGA
AGTGAACCGGCTCAAGGGCCGCGAGCCGACGCGAGTGGCCGAGCTCTACGAGGAGGAGCTGCGGGAGCTGCGGCGCCAGGTGGAGGTGCTCACTAACCAG
CGCGCGCGCGTCGACGTCGAGCGCGACAACCTGCTCGACGACCTGCAGCGGCTCAAGGCCAAGTGAGGGCCCGGCACCCCAGACTCCTCTTTCTGCGGGC
AGGGCACAGGAGGCTAGGCCTGGGGTCTGGGGTCCCGCTGTCAGCACCTGCCTTCTCCCGGGGCCCGGGACCCTCTCCTGCCCCATGTGGAGAAAGGGTC
CTCCACCTGTGTGTTTCAAGGGGCCGTGACCTCCAGGTCTCTCCCCCTGCGATCCCATCTTGCACAGGAGTTTTCTTGGGGACATAGATCAGGGGGTGGA
TATGGGAGAATTTAGGGGACCCGGTGCCCTGTGGACAGCCCCGTTAAAAAGCATTTTAAGATGCTGGGGCGATATTTATGGGGTCAGGTAGTTGATGGGC
AGAGGAAGGGCTGCAGGAGGCCCAGAGGGCAGTGTAGCCAGAGGGAGAAGGGAGGCTGATAGGAGACAGGGAAAGCAGGGCAAGGGCCCAGAGTCCAAGC
AGAGGAAGGGCTGCAGGAGGCCCAGAGGGCAGTGTAGCCAGAGGGAGAAGGGAGGCTGATAGGAGACAGGGAAAGCAGGGCAAGGGCCCAGAGTCCAAGC
AACAGCTCTCAGCTCAGCTGTGATGAGGCCCTGGGGGAGGTGGGGGGAGGGGGGAGCTTGGCCCTGGGGCCTTGCCGAGACTGTGTCTTTTTACAAGGTG
AATGGACAGGCTGGAGAAAAAGGGAGTAGGTGGGGGTCACAGCTCTCAGAGAGCTTGGGAGGACCTGACTGTAGACTTCACCAGGCTCCAAGAACGAAAA
GGGCAGCAAGTGTAGCATATTTGTTGGTCCCACTTCTGACAGGCCAAGTGAGCACAGTCACCCTCCTGCCACCAAAGTCATAAATATTAATTGAGCAGCT
ATATTGGCCAGGCTGGAGCTGGGAACCAGAAACACAGAGGTGGATAAAATAGACACAGTTTCTAACCCCAGGGAGGTCACACAGTCTGGTGGGGACATAG
ACTTCAAGGGTGTGGCTCCTGGGCAGAGATTGGGCCACTTCCTGTGCCCTCCCTGGGTGGGTGGGGCCTCTCCACTCCCTGTCTCTCCTGCCTCTACCCA
GCAGCCAGGCCCTCCCGCTCTGTCCTGGACCCACCCCCTGGTCAGCCCCCGGCCAGTCGTTTCCACTGCCAGCTTTATCACCCGCAACTGTCTGTCTTTC
TGTCTGTCCCACCCAGGCTGCAGGAGGAGATTCAGTTGAAGGAAGAAGCAGAGAACAATTTGGCTGCCTTCCGAGCGGTGAGTGCCCTTCTTTTCCCCTT
GCATGGCCTCTGGCCTTGCTCTGCCCCACCTGGGTGGCGGTGACCATGTCCTTCTCGCTTGGCCTCTCCCAGGACGTGGATGCAGCTACTCTAGCTCGCA
TTGACCTGGAGCGCAGAATTGAATCTCTCAACGAGGAGATCGCGTTCCTTAAGAAAGTGCATGAAGAGGTATACCTTGGCCCCTCTTCCTGGGGTCACTG
GGCCATGGGGAAAGCAGCCGGAAAGTGGGGTTGGGGTGAGGCTCTGGCTGGGAATAGGGGTGTGAGGGTGCTGTGTGGGCCCTGAGAGGGGACTGAAGCC
CAGTCATGCCCTACAGGAGATCCGTGAGTTGCAGGCTCAGCTTCAGGAACAGCAGGTCCAGGTGGAGATGGACATGTCTAAGCCAGACCTCACTGCCGCC
CTCAGGGACATCCGGGCTCAGTATGAGACCATCGCGGCTAAGAACATTTCTGAAGCTGAGGAGTGGTACAAGTCGAAGGTGGGTGGCCTCGCCCGGGGAC
TGGCATCTCCGTCCCCCTGAATCCCAGCTTGGATGTGCTGCCTGTGGTACCATCCATGGGAGGAGAGCCCAGAGGCTTCATGCTCCCTTGCTCATCCCTA
CCCGTGCCCTGCATCCTTCTCATTTTTGGGCCCCTTTCTCTGCCCTTAGGTGTCAGACCTGACCCAGGCAGCCAACAAGAACAACGACGCCCTGCGCCAG
GCCAAGCAGGAGATGATGGAATACCGACACCAGATCCAGTCCTACACCTGCGAGATTGACGCCCTGAAGGGCACTGTGAGTCCCTGCCCACCTGGCCAGG
CCCTGCCCCTTCCTGTCTGCAGTTCACACCCTCACTTTGTGACCTTGGGCCCATCATAGATCCTCTCTGGGCCTTCATCTACTTAAATCTACAATAGGGG
TAAAACCAGACAAGTGGATTCCAGTTGGATGCTAAGGAATCAGGGGTTCCTGGGCATCTACCTATGTGGGGACTGTGAGGCTGAATGCAATGTTCCTTTG
TATCTATTTTATTCTGAGTGTTCACATATAGACTTAATTTGAGTTCAGGGTTCAACATGGCCTGGACCTGACCATCTGGAGTTGCCTGCCAGCCCCAAAG
CTTTCTTTGGGCTGCTAGTGTCCTCTTCCCTTCCTTGACCTGGGTTCCCCCTCTCCTGCAGAACGATTCCCTGATGAGGCAGATGCGGGAATTGGAGGAC
CGATTTGCCAGTGAGGCCAGTGGCTACCAGGACAACATTGCGCGCCTGGAGGAGGAAATCCGGCACCTCAAGGATGAGATGGCCCGCCATCTGCGCGAGT
ACCAGGACCTGCTCAACGTGAAGATGGCCCTGGATGTGGAGATTGCCACCTACCGGAAGCTGCTGGAGGGAGAGGAGAGCCGGTGAGGGGCCAGGCAGGA
GCCCGAGTGGGAGGTGCGGGTGCTGGGTGGTCCATTTCTGTCCCCAGGAGGCTCGAGATTACTGATTACCTCAACAAGACCTGGAAACAATTTTTTTTT
TTTTTGAGATGGAGTTTCGCTCTTGTCGCCCAGTCTGGAGTGCAATGGCACCATCTTGGCTCACTGCAACCTCCGCCTCCTGGGTTCAAGCAATTCTCCT

דוגמא: איי CpG

- נרצה למדל את רצף הדנ"א במהלך מרקובי



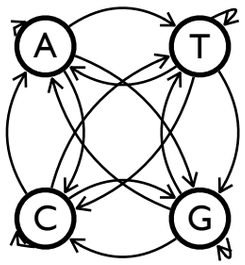
	A	C	G	T
A				
C				
G				
T				

- ובך נוכל לחשב את הניראות של הרצף

```
>chr2:220282501-220286500
GAGGCTCAGGGCTAGCTCGCCCATAGACATACATGGCAGGCAGGCTTTGGCCAGGATCCC
```

- 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 במהלך מרקובי

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

$P_-(a \to b)$

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

$P_+(a \to b)$

- והניראות עבור רצף בלשהו

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

דוגמא: איי CpG

- ובמו שעשינו בעבר, לוג יחס הנראות מתפרק לסכום

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

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]

$= \log \Bigg($

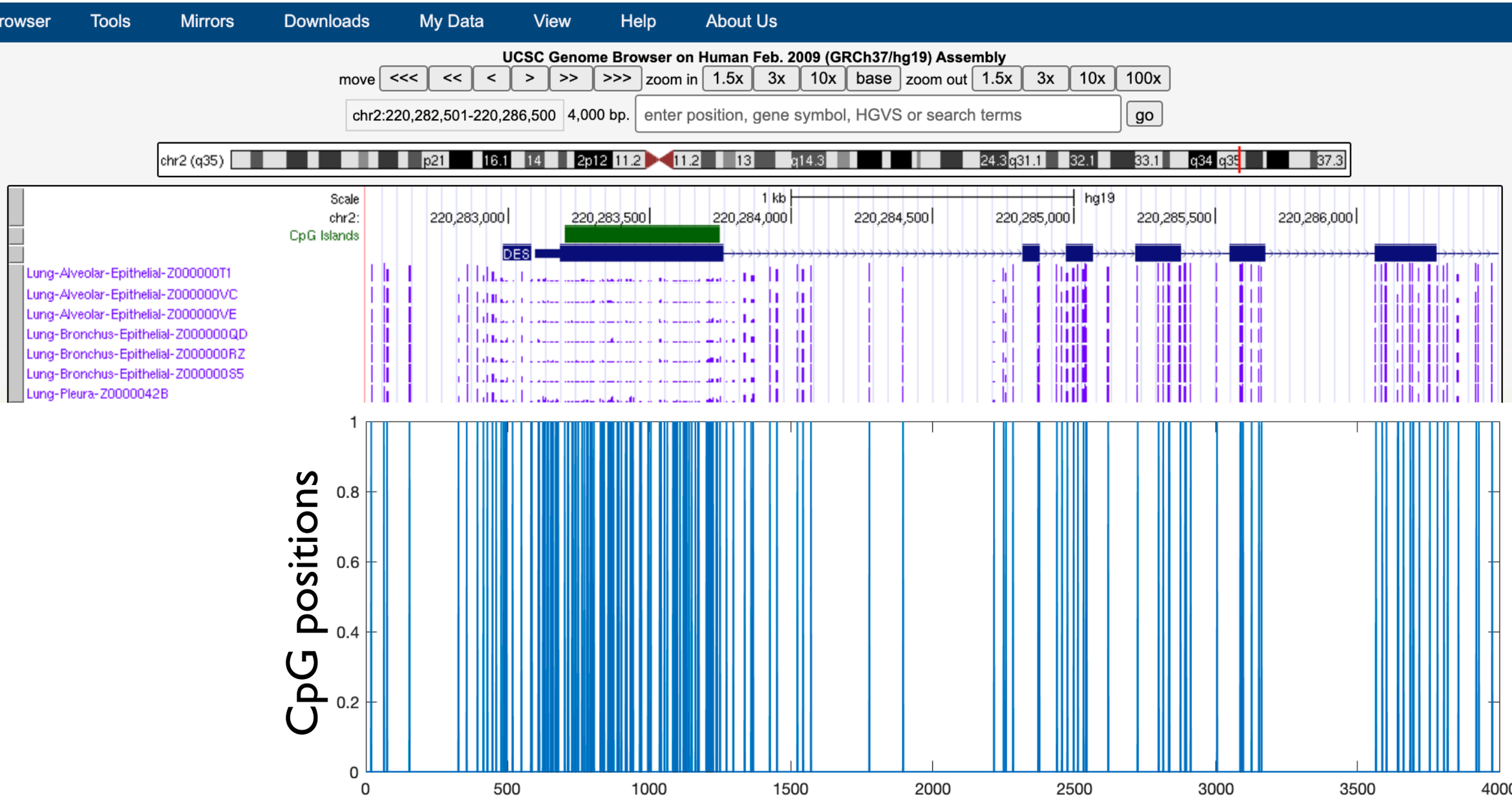
τ^{+}	A	C	G	T
A	18%	27%	43%	12%
C	17%	37%	27%	19%
G	16%	34%	38%	12%
T	8%	36%	38%	18%

$/$

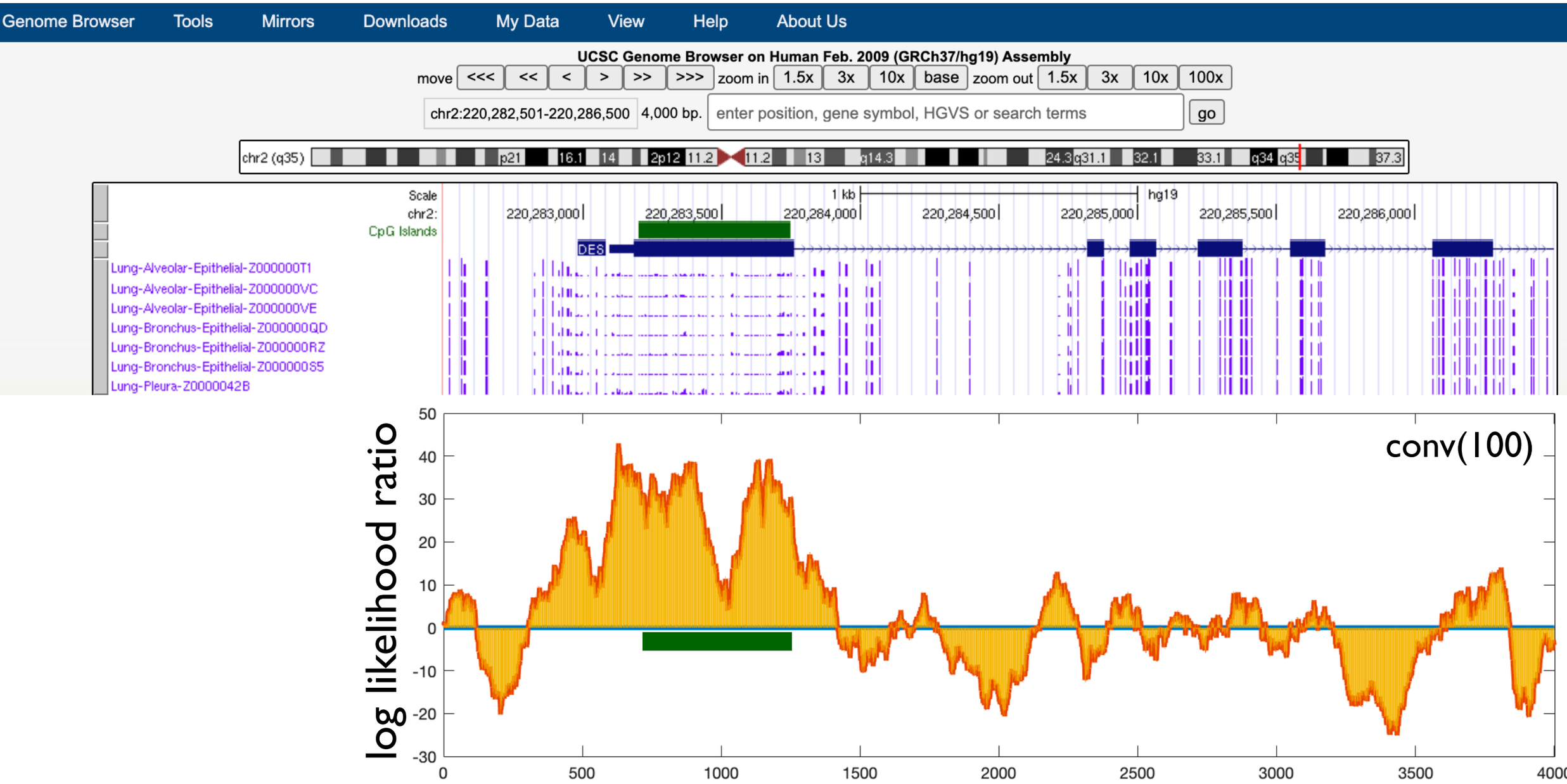
τ^{-}	A	C	G	T
A	30%	20%	29%	21%
C	32%	30%	8%	30%
G	25%	25%	30%	20%
T	18%	24%	29%	29%

$\Bigg)$

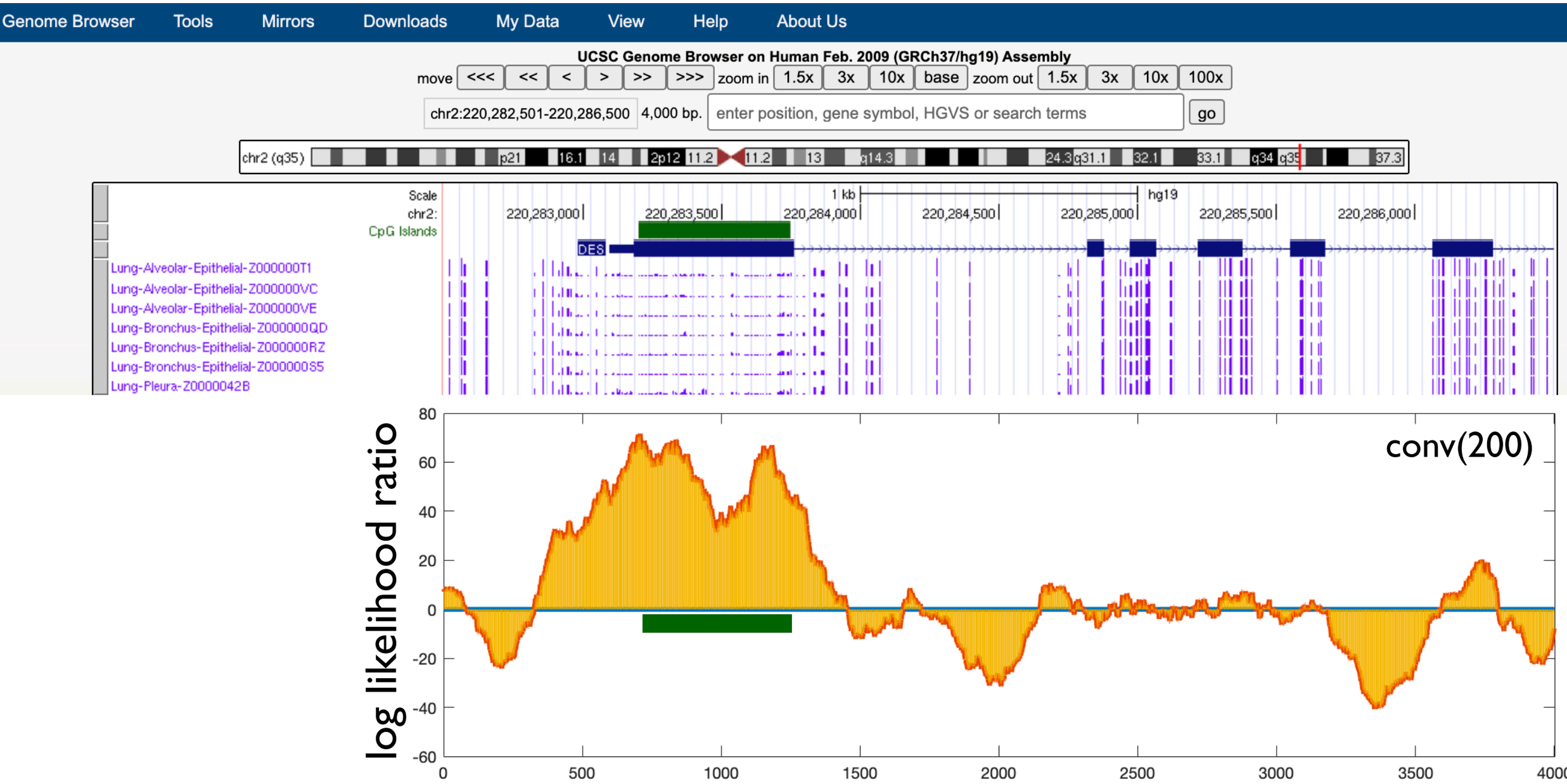
דוגמא: איי CpG



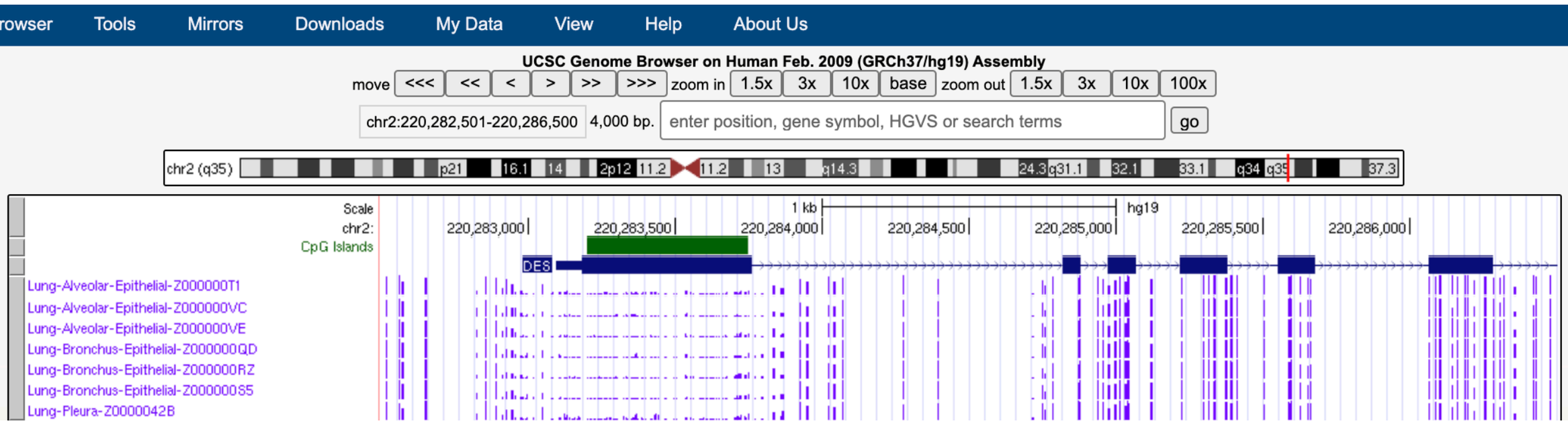
דוגמא: איי CpG

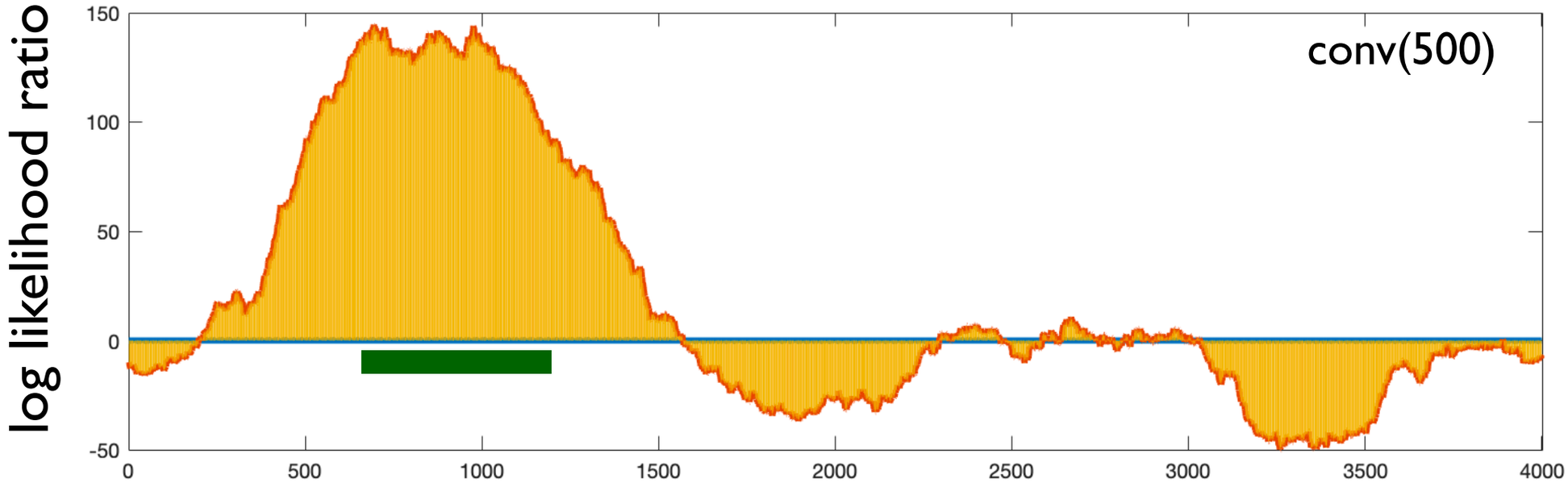


דוגמא: איי CpG



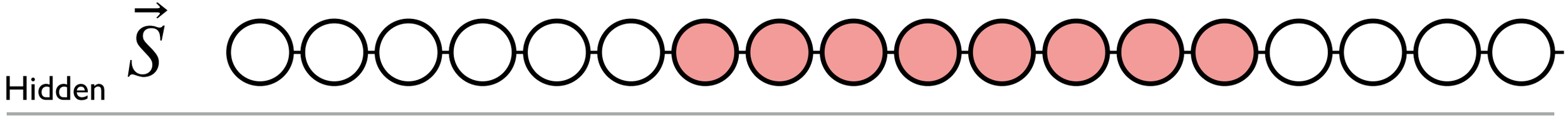
דוגמא: איי CpG



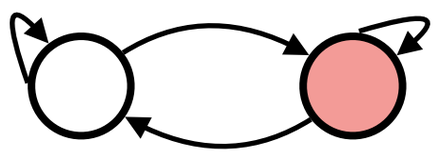


שרשראות מרקוב חבויות

- משתני התהליך המרקובי הם חבויים



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

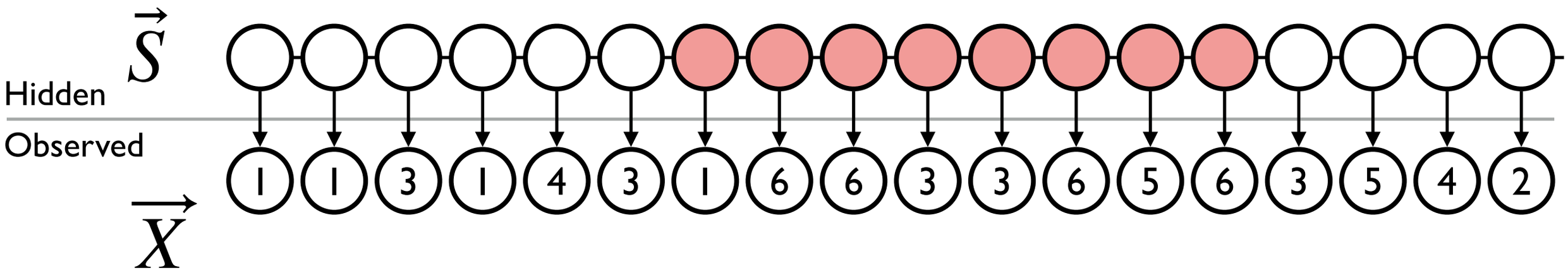


τ

τ	F	L
F	99%	1%
L	5%	95%

שרשראות מרקוב חבויות

- משתני התהליך המרקובי הם חבויים
- פולטים [בהתאם למצבם] משתנים אחרים [נצפים]

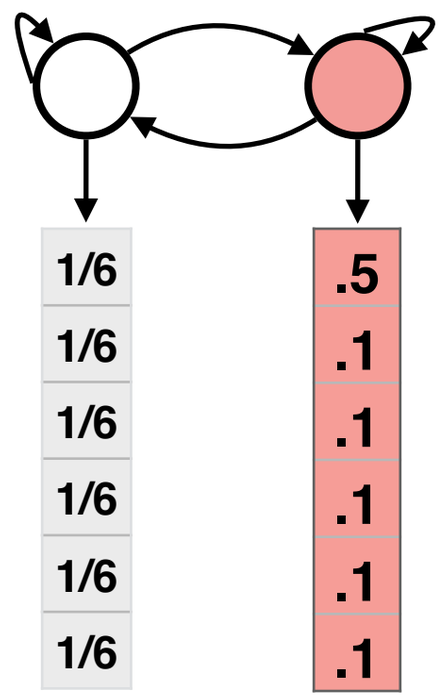


$S_i \in \{\text{Fair}, \text{Loaded}\}$

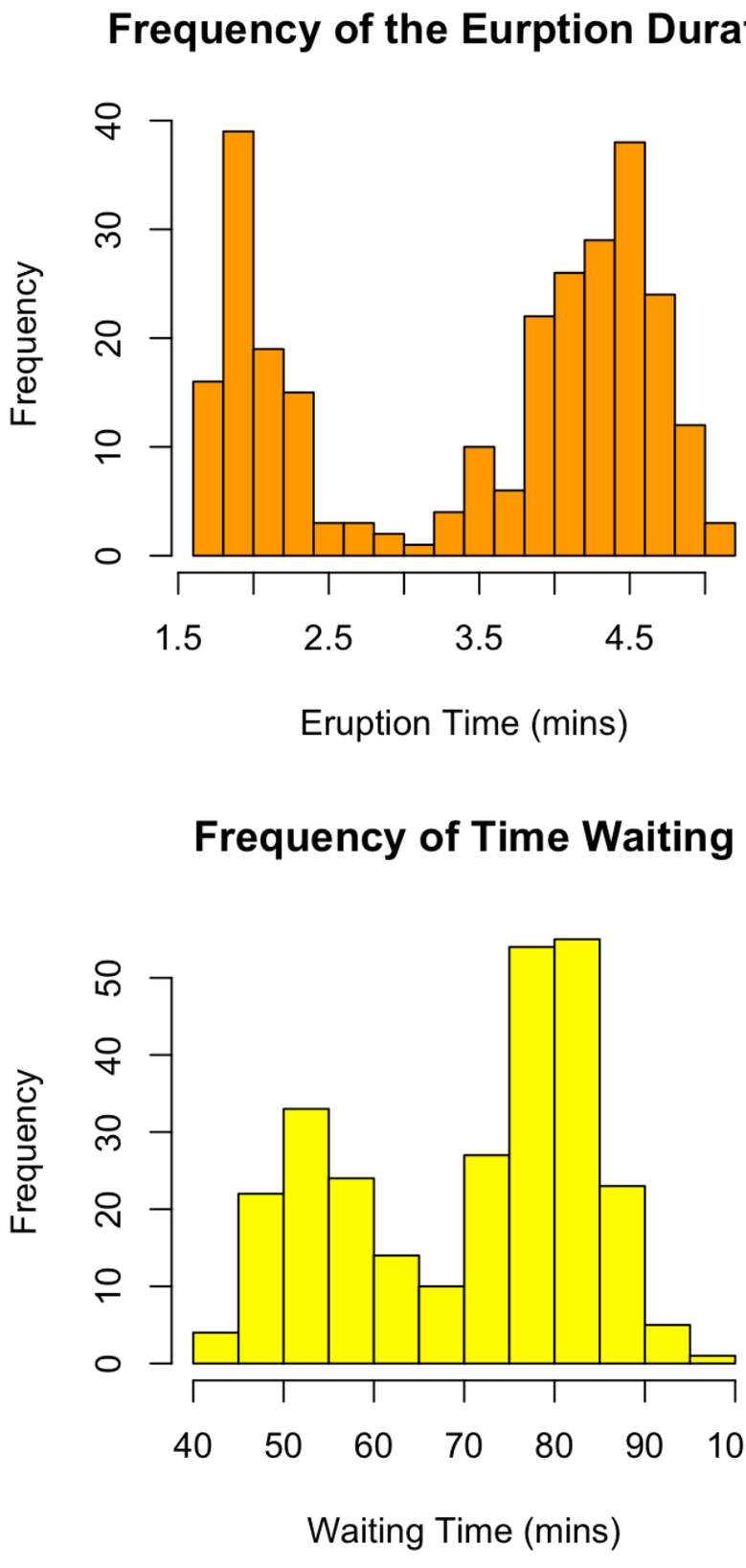
τ

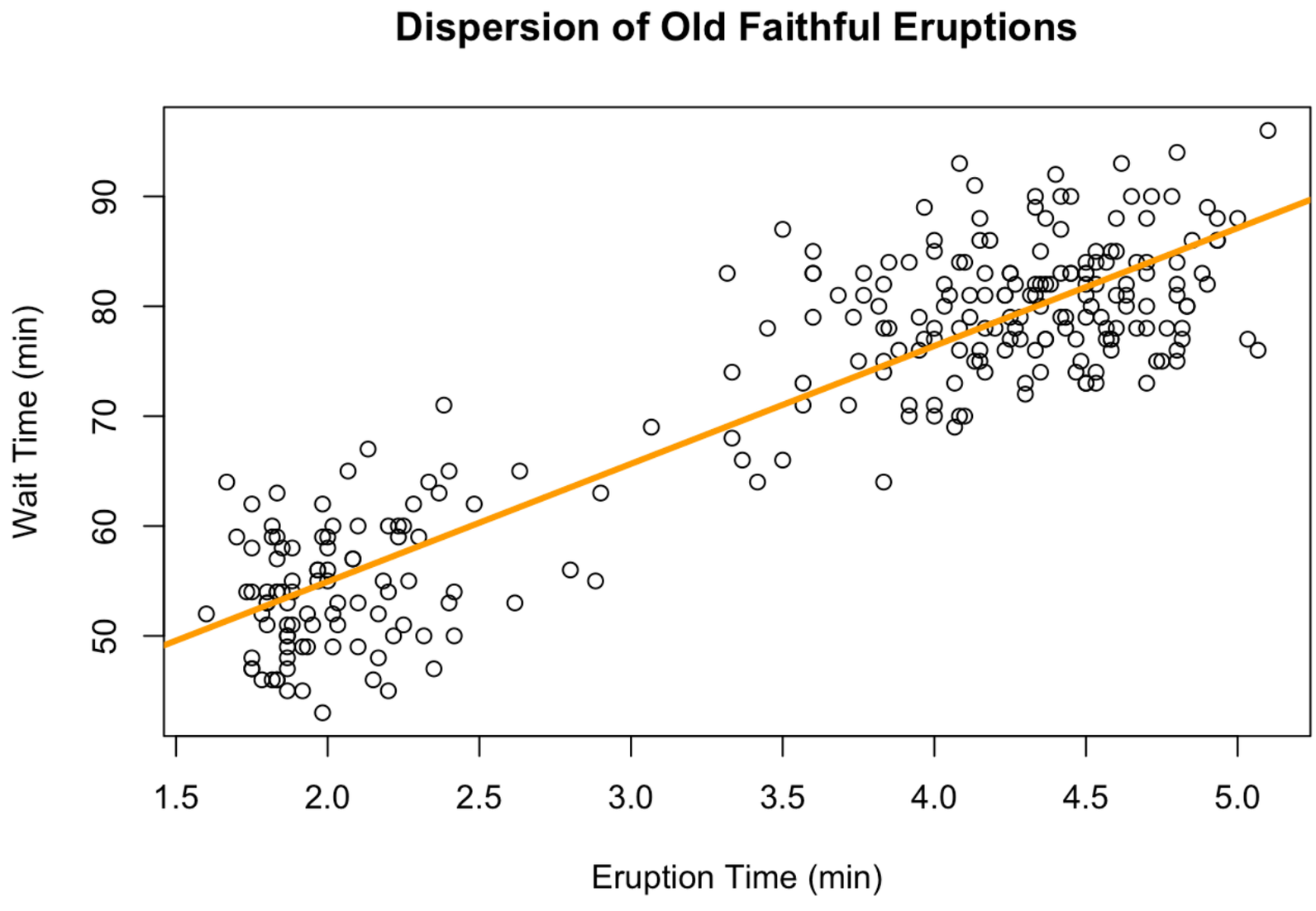
τ	F	L
F	99%	1%
L	5%	95%

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



הגייזר אולד פיית'פול

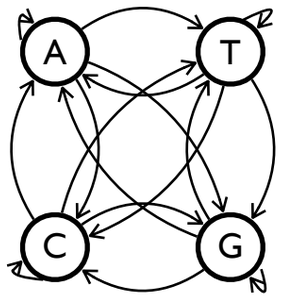




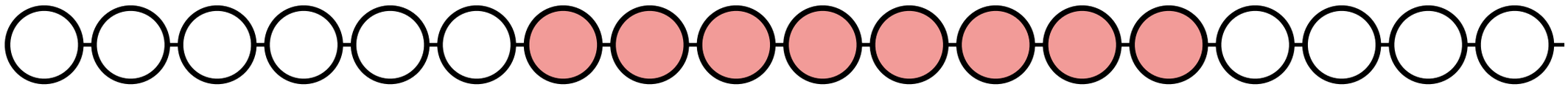
דיון - מודל מרקובי מסדר גבוה

- מודל מסדר ראשון

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

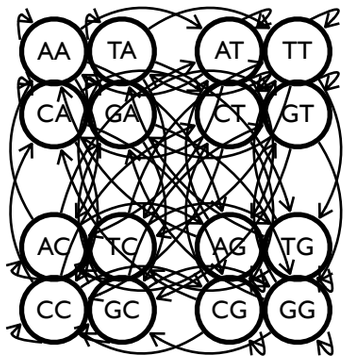


	A	C	G	T
A				
C				
G				
T				

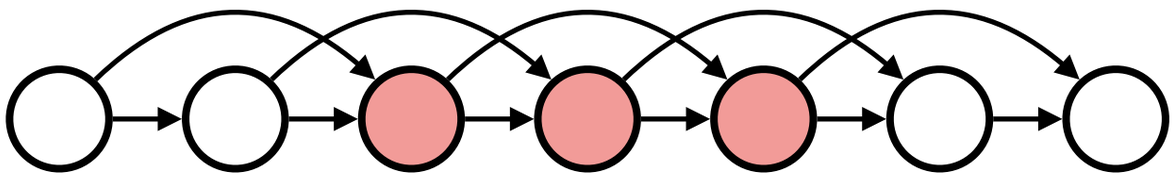


- מודל מסדר שני

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



	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.