יובל דלוס 305211880
לירון גרשוני 308350503

אלגוריתמים בביולוגיה חישובית – תרגיל 4

1 MLE for Branch Length

Suppose we are given two aligned sequences $a_1, ..., a_n$ and $b_1, ..., b_n$ (assume no gaps). In class we derived the following probabilistic model for two single character sequences:

$$P(a \xrightarrow{t} b) = P(X^{t_0+t} = b | X^{t_0} = a) = [e^{tR}]_{a,b}$$

We want to derive the MLE for branch length for the two sequences, e.g:

$$\hat{t} = \arg\max_t \prod_i \pi_{a_i} [e^{t\mathbf{R}}]_{a_i,b_i}$$

1. Define the sufficient statistics for this likelihood and derive equation for calculating the maximum-likelihood estimator $\hat{t}$. Assume independence between different positions in the alignment.

1)

1. נרצה למצוא את ערך ה-t שממקסם את ה- log-likelihood, לשם כך נתחיל מלמצוא את הלוג של הביטוי, לאחר מכן נגזור ונשווה ל-0 למציאת המקסימום.

$$\log\left(\prod_i \pi_a [e^{tR}]_{a_i,b_i}\right) = \sum_i \log\left(\pi_a [e^{tR}]_{a_i,b_i}\right) = \sum_i \log\left(\pi_{a_i}\right) + \sum_i \log\left([e^{tR}]_{a_i,b_i}\right) =$$

$$= \sum_i \log\left(\pi_{a_i}\right) + \sum_{a_i=b_i} \log\left([e^{tR}]_{a_i,b_i}\right) + \sum_{a_i\neq b_i} \log\left([e^{tR}]_{a_i,b_i}\right) =$$

$$= \sum_i \log\left(\pi_{a_i}\right) + \sum_{a_i=b_i} \log\left(\frac{1}{4}(1+3e^{-t})\right) + \sum_{a_i\neq b_i} \log\left(\frac{1}{4}(1-e^{-t})\right)$$

כעת נגדיר את הסטטיסטים בהתאם לחלקי המשוואה, כאשר נשים לב כי הביטוי האמצעי והימני קשורים זה לזה בכך שכמות האיברים שהם ביניהם היא n. נגדיר:

$$s_1 = \sum_i \log\left(\pi_{a_i}\right), \qquad s_2 = \sum_i \mathbf{1}_{[\boldsymbol{a_i}=\boldsymbol{b_i}]}$$

וכשנציב חזרה במשוואה:

$$LL = s_1 + s_2 \cdot \log\left(\frac{1}{4}(1+3e^{-t})\right) + (n-s_2)\cdot \log\left(\frac{1}{4}(1-e^{-t})\right)$$

נגזור:

$$\frac{\partial \mathrm{LL}}{\partial \mathrm{t}} = -\frac{s_2 \cdot 3e^{-t}}{1+3e^{-t}} + (n-s_2)\cdot\frac{e^{-t}}{1-e^{-t}}$$

יובל דלוס 305211880
לירון גרשוני 308350503

ונשווה ל-0:

$$\frac{s_2 \cdot 3e^{-t}}{1+3e^{-t}} = (n - s_2) \cdot \frac{e^{-t}}{1-e^{-t}}$$

$$3s_2 \cdot e^{-t}(1-e^{t}) = (n - s_2) \cdot e^{-t} \cdot (1+3e^{-t})$$

$$3s_2 \cdot e^{-t} - 3s_2 \cdot e^{-2t} = (ne^{-t} - s_2 \cdot e^{-t}) \cdot (1+3e^{-t})$$

$$3s_2 \cdot e^{-t} - 3s_2 \cdot e^{-2t} = ne^{-t} - s_2 \cdot e^{-t} + 3ne^{-2t} - 3s_2 \cdot e^{-2t}$$

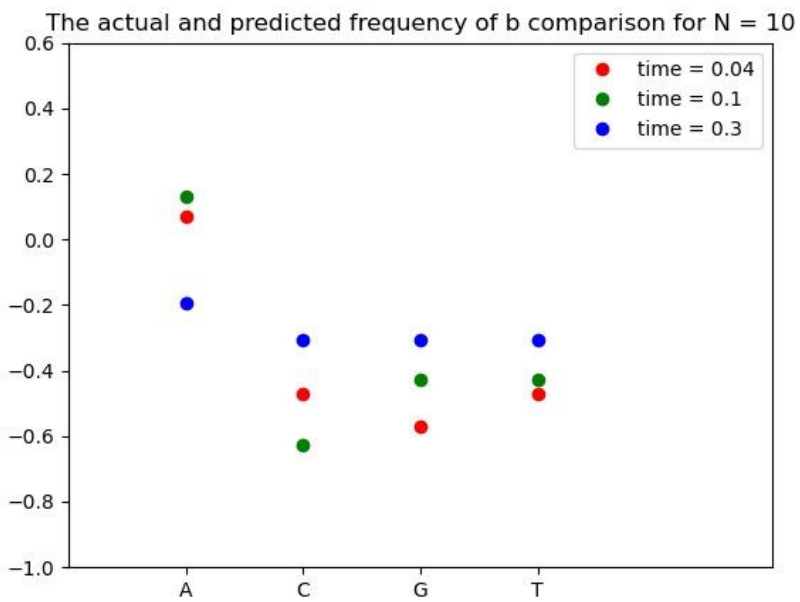
$$4s_2 \cdot e^{-t} = ne^{-t} + 3ne^{-2t} \qquad : / \, e^{-t}$$

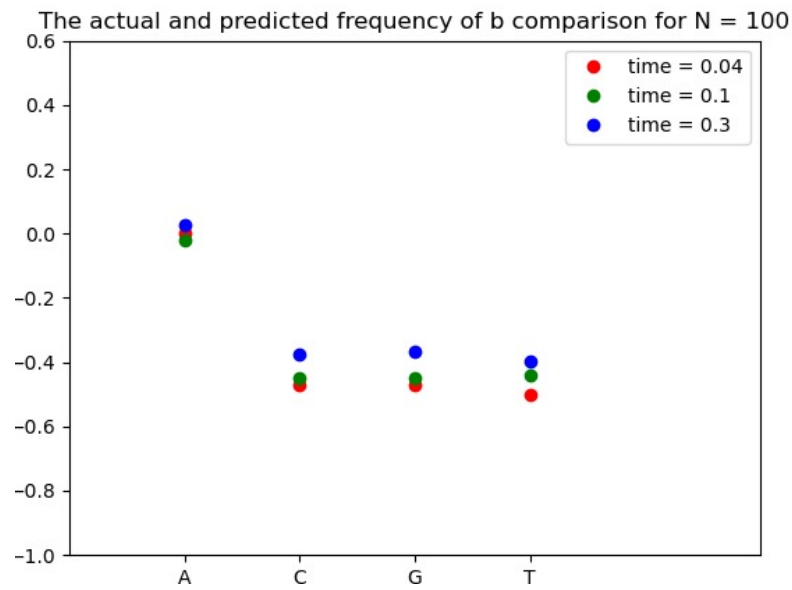
$$4s_2 = ne^{-t} + 3ne^{-t}$$

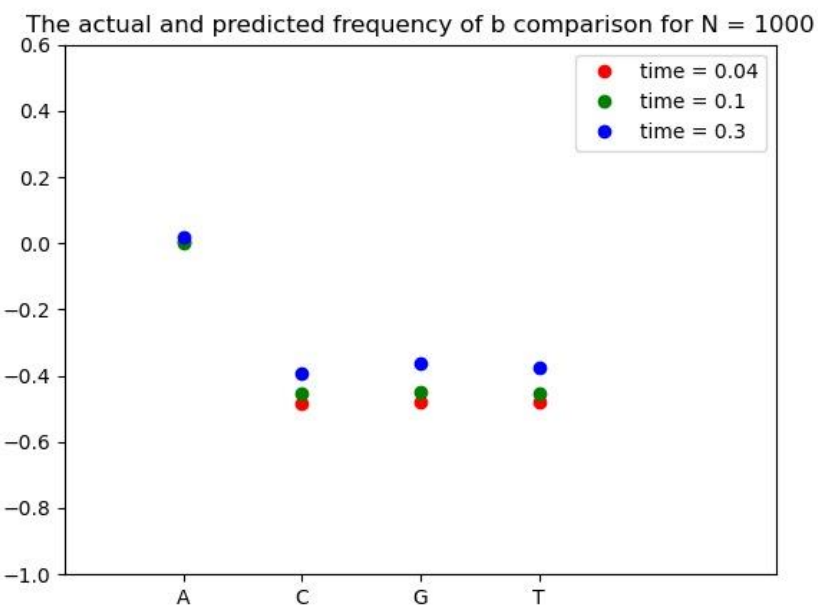
$$t = -\log\left(\frac{4s_2 - n}{3n}\right)$$

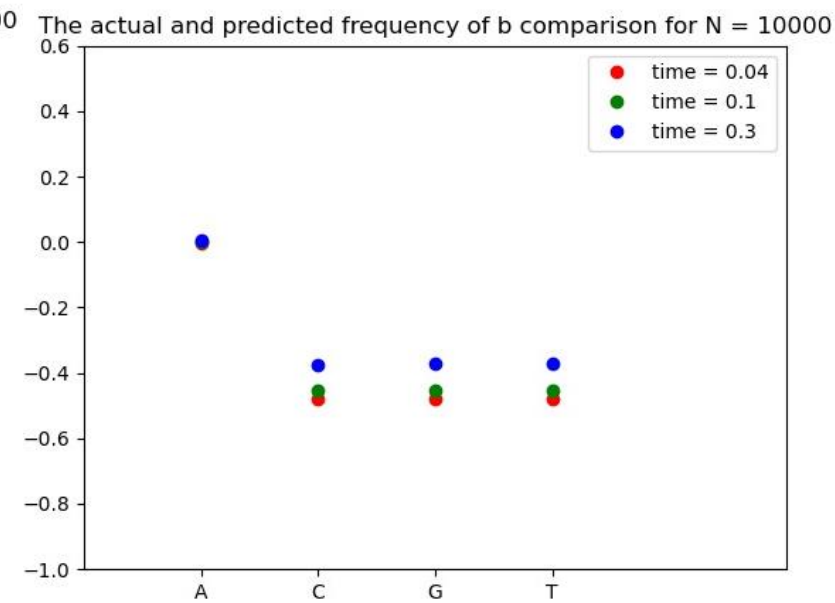
2. Building a sampler for a branch

2. נביט בגרפים הבאים:









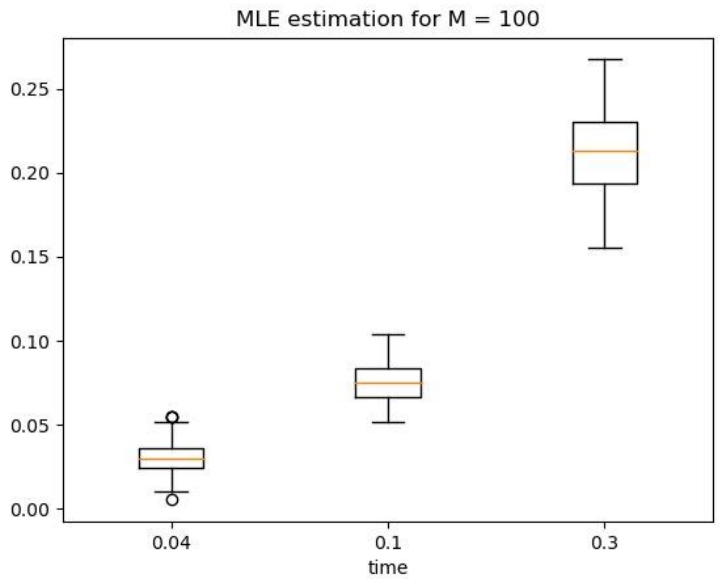
יובל דלוס 305211880
לירון גרשוני 308350503

בגרף אנו יכולים לראות את אחוזי הדיוק שהמודל שלנו פלט כאשר בחרנו כמועמד את הנוקלאוטיד A ובדקנו את הסתברויות המעבר אל הנוקלאוטידים האחרים לפי הטבלה הסטציונרית שסופקה לנו.

ניתן לראות שההבדלים בגרפים הם מספר החזרות, ובכל גרף דגמנו 3 פעמים בזמנים שונים.
נתייחס תחילה לדגימות השונות בכל גרף. ניתן לראות כי ככל שנתנו למודל לרוץ יותר זמן כך התוצאות התכנסו לערך גבוהה יותר, כלומר המודל הצליח לחזות בהצלחה את הסתברות המעבר והנוקלאוטיד החזוי.
כאשר נעבור מגרף לגרף נגדיל את מספר הדגימות עליהן המודל מתאמן, גם במקרה זה נראה שיפור של החיזויים והתכנסותם לערך יחיד.

3. Estimating the evolutionary distance

.3



ביצענו 100 הרצות שבכל הרצה הגרלנו שני רצפים באורך 500 כל אחד. עבור כל זמן t חישבנו את עומד ה-MLE, ניתן לראות שקיבלנו קירוב טוב לערכי ה-t שחזינו.
בנוסף ניתן לראות שהשונות גדלה ככל שרצים יותר זמן על הרצף, תוצאה זו הגיונית מכיוון שככל שהזמן שנותנים לרצף להשתנות ולצבור "מוטציות" כך נראה יותר שוני בין הרצף ההתחלתי לסופי. דבר זה יכול להוות בעיה אם נרצה לעשות את הפעולה ההפוכה ולשחזר את העץ למציאת אב קדמון.

2 Choosing rate matrices

1. Write a rate matrix R that converges into a **uniform** stationary distribution π but is **not reversible**.

(2

1. נרצה לבחור מטריצה שהיא לא רברסבילית, כלומר לא מקיימת את התכונה

$$\forall t,a,b \quad \pi_a P\left(a \xrightarrow{t} b\right) = \pi_b P\left(b \xrightarrow{t} a\right)$$

אבל כן מתכנסת למטריצה אחידה, נבחר במטריצה:

$$R = \begin{bmatrix} -5 & 1 & 3 & 1 \\ 3 & -5 & 1 & 1 \\ 1 & 1 & -5 & 3 \\ 1 & 3 & 1 & -5 \end{bmatrix} \overset{expm(t\cdot R)}{\underset{t\to\infty}{\Longrightarrow}} \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$

ונראה שאף שהיא מתכנסת למטריצה אחידה, היא אינה רברסבילית, למשל בכניסה

$$P(t)_{1,0} = 3 \neq 1 = P(t)_{0,1}$$

יובל דלוס 305211880
לירון גרשוני 308350503

2. Write a rate matrix R that converges into a **non-uniform** stationary distribution π but is **reversible**.

$$R = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} \overset{expm(t\cdot R)}{\underset{t\to\infty}{\Longrightarrow}} \begin{bmatrix} 4.194 & 0 & 0 & -3.194 \\ 0 & 4.194 & -3.194 & 0 \\ 0 & -3.194 & 4.194 & 0 \\ -3.194 & 0 & 0 & 4.194 \end{bmatrix}$$

זו מטריצה די שטותית, שמאפשרת לכל נוקלאוטיד להשתנות לאחד אחר בלבד, אך היא רברסביבילית, וכפי שראינו אינה מתכנסת לאחידה.

3 Probabilistic Model of Evolution

Suppose we are given a binary tree T with leaves labeled 1...n and n-1 internal nodes. Assume 2n-1 is the root of the tree and let $\tau = \{t_{ij} | (i \overset{t_{ij}}{\to} j) \in T\}$ be the set of branch lengths.

Let $X_1, X_2, ...X_n, ...X_{2n-1}$ be random variables representing the sequences in the tree. For simplicity, let's assume that all the sequences are of length 1.

We would like to show that if R generates reversible matrices P, then we are allowed to reroot the tree without changing the likelihood.

We can calculate the likelihood using:

$P(X_1, X_2, ...X_{2n-1}) = P(X_{2n-1}) \prod_{(i\to j)\in T} P(X_i \overset{t_{ij}}{\to} X_j)$

(We sample the root and then sample each node given its parent)

where $P(a \overset{t}{\to} b) = [e^{tR}]_{a,b}$ and $P(X_{2n-1} = a) = \pi_a$ is the stationary distribution.

1. Please show that the likelihood can also be written as:

$$P(X_1, X_2, ...X_{2n-1}) = \left[\prod_i \pi_{X_i}\right] \prod_{(i\to j)\in T} \frac{[e^{t_{ij}R}]_{X_iX_j}}{\pi_{X_j}}$$

(3

1. נפתח את המשוואה:

$$P(X_1, X_2, \dots, X_{2n-1}) = P(X_{2n-1}) \prod_{(i\to j)\in T} P\left(X_i \overset{t_{i,j}}{\to} X_j\right) \overset{(*)}{=}$$

$$= \pi_{X_{2n-1}} \prod_{(i\to j)\in T} [e^{t_{ijR}}]_{X_i,X_j} \overset{\frac{\pi_{X_j}}{\pi_{X_j}}}{=}$$

$$= \pi_{X_{2n-1}} \prod_{(i\to j)\in T} \frac{\pi_{X_j}[e^{t_{ijR}}]_{X_i,X_j}}{\pi_{X_j}} =$$

$$= \pi_{X_{2n-1}} \prod_{(i\to j)\in T} \pi_{X_j} \prod_{(i\to j)\in T} \frac{[e^{t_{ijR}}]_{X_i,X_j}}{\pi_{X_j}} \overset{(**)}{=}$$

$$= \pi_{X_{2n-1}} \prod_{\boldsymbol{i=1}}^{\boldsymbol{2n-2}} \pi_{X_j} \prod_{(i\to j)\in T} \frac{[e^{t_{ijR}}]_{X_i,X_j}}{\pi_{X_j}} =$$

יובל דלוס 305211880
לירון גרשוני 308350503

$$= \prod_{i=1}^{2n-1} \pi_{X_j} \prod_{(i\to j)\in T} \frac{[e^{t_{ijR}}]_{X_i,X_j}}{\pi_{X_j}}$$

כאשר:

$(*)$ נכתוב את הביטויים לפי הגדרה.

$(**)$ נתון לנו שזהו עץ בינארי , לכן יש לו $2n-2$ צלעות, כמספר הקודקודים פחות קודקוד השורש אליו אין אף צלע שנכנסת לכן לא נספור אותו.

2. Now assume the underlying Markov process is reversible, and show that re-rooting the tree will not change the joint distribution.

2. נרצה לקחת את העץ "ולשתול" אותו מחדש עם שורש אחר, נעשה זאת ונשתמש בתוצאה מהסעיף הקודם לפתוח אותו בהתאם, נבחר את השורש החדש להיות X_{2n-2}:

$$P(X_1, X_2, \dots, X_{2n-1}) = P(X_{2n-2}) \prod_{(i\to j)\in T} P\left(X_i \xrightarrow{t_{i,j}} X_j\right) \overset{(*)}{=}$$

$$= P(X_{2n-2}) \prod_{\substack{(i\to j)\in T \\ i\neq 2n-2 \\ j\neq 2n-1}} P\left(X_i \xrightarrow{t_{i,j}} X_j\right) \cdot P\left(X_{2n-2} \xrightarrow{t_{(2n-2),(2n-1)}} X_{2n-1}\right) \overset{(**)}{=}$$

$$= P(X_{NR}) \prod_{\substack{(i\to j)\in T \\ i\neq NR \\ j\neq OR}} P\left(X_i \xrightarrow{t_{i,j}} X_j\right) \cdot P\left(X_{NR} \xrightarrow{t_{NR,OR}} X_{OR}\right) \overset{(***)}{=}$$

$$= P(X_{NR}) \prod_{\substack{(i\to j)\in T \\ i\neq NR \\ j\neq OR}} P\left(X_i \xrightarrow{t_{i,j}} X_j\right) \cdot \frac{\pi_{OR} \cdot P\left(X_{OR} \xrightarrow{t_{OR,NR}} X_{NR}\right)}{\pi_{NR}} =$$

$$= \pi_{NR} \prod_{\substack{(i\to j)\in T \\ i\neq NR \\ j\neq OR}} [e^{t_{ijR}}]_{X_i,X_j} \cdot \frac{\pi_{OR} \cdot [e^{t_{OR,NR}R}]_{X_{OR},X_{NR}}}{\pi_{NR}} =$$

$$= \boldsymbol{\pi_{OR}} \prod_{\substack{(i\to j)\in T \\ i\neq NR \\ j\neq OR}} [e^{t_{ijR}}]_{X_i,X_j} \cdot \frac{\boldsymbol{\pi_{NR}} \cdot [e^{t_{NR,OR}R}]_{X_{NR},X_{OR}}}{\pi_{NR}} =$$

$$= \pi_{OR} \prod_{(i\to j)\in T} [e^{t_{ijR}}]_{X_i,X_j} = P(X_{2n-1}) \prod_{(i\to j)\in T} P\left(X_i \xrightarrow{t_{i,j}} X_j\right)$$

כנדרש.

כאשר:

$(*)$ נוציא באופן מפורש את המעבר מהשורש החדש לשורש הקודם.

$(**)$ לצורך נוחות כתיבה נסמן:

יובל דלוס 305211880
לירון גרשוני 308350503

$$OR \coloneqq Original\ Root := 2n - 1$$

$$NR \coloneqq New\ Root := 2n - 2$$

(***) השתמשנו בתכונת הרברסיביליות שלמדנו בשיעור:

$$\pi_a \cdot P(t)_{a,b} = \pi_b \cdot P(t)_{b,a} \Longrightarrow$$

$$\Longrightarrow P(t)_{a,b} = \frac{\pi_b}{\pi_a} \cdot P(t)_{b,a}$$