المجلة العراقية للعلوم الإحصائية

http://stats.uomosul.edu.iq

# تقدير مصفوفة الاحتمالات الانتقالية لسلسلة جين E.Coli بطريقة ذكائية مقترحة وطرائق تقليدية

**علاء فرحان احمد عزاوي[1] مثنى صبحي سليمان[2]**

[1،2]قسم الإحصاء والمعلوماتية، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل ، العراق

**معلومات النشر**

**تاريخ المقالة:**
**تاريخ الاستلام:18 كانون الاول 2023**
**تاريخ القبول: 20 أيار 2023**
**تاريخ القبول : 26 ايار 2023**
**متاح على الانترنيت 1 حزيران 2024**

**الكلمات الدالة:**
**احتمالات الانتقال، سلاسل ماركوف، الإمكان الأعظم، بيز، DNA**

**المراسلة:**
**علاء فرحان احمد**
alaa.21csp99@student.uomosul.edu.iq

**المستخلص**

في كثير من الاحيان تكون مقدرات المصفوفة الانتقالية لسلسلة ماركوف غير دقيقة ويتم اعتبار المصفوفة الانتقالية معطاة، توجد العديد من الطرق التي تستخدم لتقدير مصفوفة الاحتمالات الانتقالية للحالات المختلفة اشهرها هي طريقة الإمكان الأعظم، ومن اجل العثور على مقدر جيد لمصفوفة الاحتمالات الانتقالية لسلسلة ماركوف تم استخدام طريقة بيز وطريقة مقترحة في هذا البحث وذلك للوصول الى احتمالات انتقالية باقل تباين، وقد تم اختيار سلسلة جين بكتيريا القولون الاشريكية ( E.Coli ) Escherichia Coli كجانب تطبيقي للدراسة وذلك لاهميتها في الأبحاث الطبية ولغرض اكتشاف وتصنيع العلاجات من خلال معرفة الشكل النهائي لسلسلة الجين الخاص بها، وبعد اختبار سلسلة جين E.Coli تبين ان هذه السلسلة تمثل سلسلة ماركوف، و تم تقدير كل من مصفوفة الاحتمالات الانتقالية وتباين الاحتمالات الانتقالية باستخدام طريقة بيز والامكان الأعظم والطريقة المقترحة، وتبين ان مقدر الطريقة المقترحة للاحتمالات الانتقالية افضل من طريقة الإمكان الأعظم وطريقة بيز بالاعتماد على التباين.

DOI 10.33899/IQJOSS.2024.183250 , ©Authors, 2024, College of Computer Science and Mathematics University of Mosul.
This is an open access article under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

## 1. مقدمة Introduction :

تعد سلاسل ماركوف Markov Chain من اهم العمليات التصادفية حيث قام العالم اندريه ماركوف بنشر مجموعة من الأوراق العلمية عام 1907، مثلت هذه الأوراق تطور لعملية كالتون–واتسون حيث تعتبر هي البداية لسلاسل ماركوف. ومن خلال هذه الأوراق التي قدمها ماركوف وضع شرط لاستخدام سلاسل ماركوف يمثل هذا الشرط الخاصية الماركوفية والتي تنص: " ان الأوضاع المستقبلية للعملية تعتمد فقط على الوضع الحالي للعملية ومستقلة عن الأوضاع السابقة "(Yang & Sha, 2011))وتوجد أربع حالات لسلاسل ماركوف كما هو في العمليات التصادفية حيث يمكن ان يكون كل من الزمن وفضاء الحالة مستمر او متقطع، ومن أكثر سلاسل ماركوف تطبيقاً وأشهرها هي التي يكون فيها كل من فضاء الحالة والزمن متقطع.

لذا يمكن ترميز عناصر فضاء الحالة بالأعداد الصحيحة $(1,2,3,\dots)$. فاذا كان المتغير العشوائي المتقطع $X_n$ يمثل حالة ما مشاهدة عند الزمن $n$ فيمكن ترميز العملية التصادفية والتي تكون بفضاء حالة متقطع بالشكل $\{X_n ; n = 0,1,2,3, \dots\}$، حيث ان الدليل $n$ يرمز للزمن او الموقع في مكان ما، ويمكن ان يكون أي دليل اخر. وبما ان الدليل $n$ يمثل الزمن فأنه يمثل الحاضر، و $n - c$ يمثل الماضي قبل $c$ من وحداة الزمن و $n + c$ يمثل المستقبل بعد $c$ من وحداة الزمن.

وتسمى العملية التصادفية $X_n$ بعملية ماركوف اذ حققت الاتي:

$$P_r\{ X_n = j \,|X_{n-1} = i \, , X_{n-2} = c \, , \dots , X_k = k \, \} = P_r\{ \, X_n = j \mid X_{n-1} = i \, \} \quad (1)$$

حيث يوضح هذا الاحتمال الشرطي بأن العملية في الحالة ( $i$ ) سوف تنتقل للحالة ( $j$ ) بعد خطوة واحدة وباحتمال $P_{ij}$، وقام الباحث G. Wang بتاريخ 2010 بنشر بحث حول تقدير الاحتمالات الانتقالية(*Wang, 2010*)، وقام الباحثون , Junsheng Ma واخرون في عام 2014 بنشر بحث حول أسلوب بيز في تقدير مصفوفة الاحتمالات الانتقالية للبيانات ذات الزمن المتقطع لعملية ماركوف ((Ma, Yu, Symanski, Doody, & Chan, 2014)، وقام ايضاً Lee واخرون في عام 1968 بنشر بحث حول مقدر بيز والامكان الأعظم للاحتمالات الانتقالية(*Lee, Judge, & Zellner, 1968*)، وقدم ايضاً كل من KALBFLEISCH و LAWLESS في عام 1984 تقدير المربعات الصغرى للاحتمالات الانتقالية من تجميع البيانات((Kalbfleisch & Lawless, 1984).

**2. الجانب النظري**

بناء نموذج لسلسلة ماركوف من خلال افتراض سلسلة من المشاهدات فأن اول ما يتم ملاحظته هو مشاهدات تلك السلسلة، والتي قد تكون أحرف او اعداد صحيحة. ومن هذه المشاهدات يتم حساب الانتقالات من حالة الى أخرى لتكوين مصفوفة تضم هذه الانتقالات تسمى بمصفوفة التكرارات ويرمز لها بالرمز $F$. وعلى فرض ان فضاء الحالة $s = \{ \, 1, 2, 3, \dots , n \, \}$ وبذلك تكون مصفوفة التكرارات بالشكل التالي:

$$F = \begin{matrix} \\ 1 \\ 2 \\ \vdots \\ n \end{matrix} \begin{matrix} \begin{matrix} 1 & 2 & 3 & \cdots & n \end{matrix} \\ \begin{bmatrix} f_{11} & f_{12} & f_{13} & \cdots & f_{1n} \\ f_{21} & f_{22} & f_{23} & \cdots & f_{2n} \\ \vdots & & \vdots & \cdots & \vdots \\ f_{n1} & f_{n2} & f_{n3} & \cdots & f_{nn} \end{bmatrix} \end{matrix} \quad (2)$$

بما ان فضاء الحالة متقطع ومنتهي بالإمكان التعبير عن احتمالات الانتقال من والى الحالات المختلفة بعد خطوة واحدة بشكل مصفوفة تسمى بالمصفوفة الانتقالية *Transition Matrix* ويرمز لها بالرمز $P$، ويمثل العنصر $(i, j)$ من هذه المصفوفة باحتمال الانتقال من الحالة $i$ الى الحالة $j$ باحتمالية $P_{ij} = (X_n = j | X_{n-1} = i)$ . كما ان المصفوفة الانتقالية $P$ هي عبارة عن مصفوفة تصادفية *Stochastic Matrix* ويجب ان تحقق الشرطين:

1- جميع عناصرها غير سالبة واكبر من الصفر ( كونها قيم احتمالية).

2- مجموع كل صف من صفوفها يساوي الواحد ( كون مجموع الاحتمالات الكلية يساوي الواحد).

فاذا كان فضاء الحالة $S = \{1,2, \dots , n\}$، فأن الشكل العام للمصفوفة الانتقالية ذات الخطوة الواحدة يكون بالشكل:

$$P = \begin{matrix} X_n \\ X_{n-1} \\ 1 \\ \vdots \\ N \end{matrix} \begin{matrix} \begin{matrix} 1 & 2 & \dots & N \end{matrix} \\ \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ \vdots & & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \end{matrix} \quad (3)$$

**3. تقدير مصفوفة الاحتمالات الانتقالية Estimation of Probability Transition Matrix :**

هناك العديد من الطرق التي تستخدم لتقدير مصفوفة الاحتمالات الانتقالية للحالات المختلفة، وسيتم تناول طريقة الإمكان الأعظم وطريقة بيز وطريقة مقترحة لتقدير مصفوفة الاحتمالات الانتقالية.

**1.3 طريقة الإمكان الأعظم Maximum Likelihood Method (MLE) :**

ليكن $x_1, x_2, x_3, \dots, x_n$ هي مشاهدات عينة عشوائية مسحوبة من مجتمع بدالة كثافة احتمالية $f(x, P)$. عندئذ يمكن تعريف دالة الامكان لمشاهدات العينة بانها التوزيع المشترك لتلك المشاهدات. ليكن $L(P)$ هو رمز لدالة الامكان فان هذه الدالة تكون بالشكل( Guo, Liao, Zhao, & Mettas, 2007; Singer, Helic, Taraghi, & Strohmaier, 2014):

$$L(P) = \prod_{i=1}^{n} f(x_i, P) \quad (4)$$

$$L(P) = P(x_1)P(x_2|x_1) \dots P(x_{n-1}|x_{n-2})P(x_n|x_{n-1})$$

$$L(P) = P(x_1)\prod_{t=2}^{n} P(x_t|x_{t-1})$$

$$L(P) = P(x_1)\prod_{t=2}^{n} P_{x_{t-1}x_t} \quad (5)$$

وبعد اعادة كتابة احتمالات الانتقال للحصول على دالة الامكان لمصفوفة الانتقال:

$$L(P) = P(x_1)\prod_{i=1}^{k}\prod_{j=1}^{k} P_{ij}^{n_{ij}} \quad (6)$$

حيث ان $n_{ij}$ عدد الانتقالات من $i$ الى $j$

ومن خلال تعظيم المعادلة أعلاه بأخذ اللوغارتم الطبيعي نحصل على :

$$\ln L(P) = \ln P(x_1) + \sum_{i,j} n_{ij} \ln P_{ij} \quad (7)$$

ولغرض إيجاد احتمالات الانتقال من معادلة الإمكان الأعظم اعلاه نتبع طريقة مضاعفات لاكرانج

$$\mathcal{L}(P, \lambda) = \ln P(x_1) + \sum_{i,j} n_{ij} \ln P_{ij} - \left(\sum_{i=1}^{j} \lambda_i \left(\sum_j P_{ij} - 1\right)\right) \quad (8)$$

$\lambda_{ij}$ : مضاعف لاكرانج ، $P_{ij}$ : الاحتمالات الانتقالية ، $\sum_j P_{ij} = 1$ ، وبأخذ الاشتقاق الجزئي للمعادلة (8) بالنسبة لـ $P_{ij}$ ولـ $\lambda_i$ نحصل على:

$$\frac{\partial \mathcal{L}}{\partial \lambda_i}\left[\ln P(x_1) + \sum_{i,j} n_{ij} \ln P_{ij} - \left(\sum_{i=1}^{j} \lambda_i \left(\sum_j P_{ij} - 1\right)\right)\right] = \sum_j P_{ij} - 1 \quad (9)$$

وبما ان

$$\sum_j P_{ij} = 1$$

أي ان

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = 0$$

$$\frac{\partial \mathcal{L}}{\partial P_{ij}}\left[\ln P(x_1) + \sum_{i,j} n_{ij} \ln P_{ij} - \left(\sum_{i=1}^{j} \lambda_i \left(\sum_j P_{ij} - 1\right)\right)\right] = \frac{n_{ij}}{P_{ij}} - \lambda_i \quad (10)$$

وبمساواة المعادلة (10) بالصفر نحصل على:

$$\frac{n_{ij}}{P_{ij}} - \lambda_i = 0$$

$$\lambda_i = \frac{n_{ij}}{P_{ij}}$$

$$P_{ij} = \frac{n_{ij}}{\lambda_i} \quad (11)$$

وبأخذ المجموع لطرفي المعادلة اعلاه الى ( $n$ ) بالنسبة لـ ( $j$ ) نحصل:

$$1 = \sum_j \frac{n_{ij}}{\lambda_i}$$

$$\lambda_i = \sum_j n_{ij} \quad (12)$$

وبتعويض المعادلة (12) في المعادلة (11) نحصل على:

$$P_{ij}^{(MLE)} = \frac{n_{ij}}{\sum_j n_{ij}} \quad (13)$$

وعلية فأن الصيغة (13) تمثل مقدر لمصفوفة الاحتمالات الانتقالية لسلسلة ماركوف باستخدام طريقة الإمكان الأعظم.

**2.3 طريقة بيز Bayes method :**

ليكن لدينا $D = (x_1, x_2, \dots, x_n)$ تمثل البيانات المشاهدة لذا يمكن كتابة نظرية بيز بالشكل التالي:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \quad (14)$$

حيث ان $P(\theta|D)$ هو التوزيع اللاحق والتي تمثل احتمالية معلمات النموذج $(\theta)$ مشروطة بالبيانات المشاهدة . والمقدار $P(D|\theta)$ الاحتمال الشرطي للبيانات بالنظر لمعلمات النموذج، ويمثل هذا المقدار دالة الإمكان . بينما المقدار $P(D)$ احتمال بيانات النموذج . والمقدار $P(\theta)$ يمثل الاحتمال الاولي ويعكس الاحتمال الاولي اعتقادنا حول المعلمات قبل ان نرى البيانات، في حالتنا هذه الاحتمال الاولي لكل صف من مصفوفة الانتقال هو توزيع Dirichlet distribution.

ليكن لدينا المتغيرات $(x_1, x_2, \dots, x_k)$ وان هذه المتغيرات يجب ان تحقق $x_i \in [0,1]$ و $\sum_{i=1}^{k} x_i = 1$ ، ومعلمات التوزيع $(\alpha_1, \alpha_2, \dots, \alpha_k)$، $\alpha_0 = \sum_{i=1}^{k} \alpha_i$ ، لذلك فان توزيع Dirichlet distribution يكون بالشكل( Demirel & Çelik, (2018))(Lin, 2016):

$$f(x_1, x_2, \dots, x_k; \alpha_1, \alpha_2, \dots, \alpha_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1} \quad (15)$$

ان القيمة المتوقعة لتوزيع Dirichlet distribution هي:

$$E(x_i) = \frac{\alpha_i}{\alpha_0} \quad (16)$$

كما ان تباين توزيع Dirichlet distribution هو:

$$Var(x_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} \quad (17)$$

حسب طريقة بيز لتقدير مصفوفة الاحتمالات الانتقالية، فأن الاحتمال الاولي للمصفوفة هو نتيجة توزيع Dirichlet distribution لكل صف والذي يكون بالشكل:

$$P(\theta) = \prod_i \frac{\Gamma(\sum_j \alpha_{ij})}{\prod_j \Gamma(\alpha_{ij})} \prod_j P_{ij}^{\alpha_{ij} - 1} \quad (18)$$

وان الاحتمال الشرطي للبيانات بالنظر لمعلمات النموذج يكون بالشكل:

$$P(D|\theta) = P(x_1) \prod_i \prod_j P_{ij}^{n_{ij}} \quad (19)$$

وان احتمال بيانات النموذج يكون بالشكل:

$$P(D) = \int P(D|\theta)\, P(\theta)\, d\theta \qquad (20)$$

وبتعويض المعادلة (18) والمعادلة (19) في المعادلة (20) نحصل على:

$$P(D) = \int P(x_1) \prod_i \prod_j P_{ij}^{n_{ij}} \prod_i \frac{\Gamma(\sum_j \alpha_{ij})}{\prod_j \Gamma(\alpha_{ij})} \prod_j P_{ij}^{\alpha_{ij}-1} d\theta$$

$$\because \int \prod_j P_{ij}^{n_{ij}+\alpha_{ij}-1}\, d\theta = \frac{\prod_j \Gamma(n_{ij}+\alpha_{ij})}{\Gamma(\sum_j (n_{ij}+\alpha_{ij}))}$$

$$\therefore P(D) = P(x_1) \prod_i \frac{\Gamma(\sum_j \alpha_{ij}) \prod_j \Gamma(n_{ij}+\alpha_{ij})}{\prod_j \Gamma(\alpha_{ij}) \Gamma(\sum_j (n_{ij}+\alpha_{ij}))} \qquad (21)$$

يتم الحصول على التوزيع اللاحق لطريقة بيز في تقدير مصفوفة الاحتمالات الانتقالية من خلال تعويض المعادلة (18) والمعادلة (19) والمعادلة (21) في المعادلة (14)

$$P(\theta|D) = \prod_i \prod_j P_{ij}^{n_{ij}+\alpha_{ij}-1} \frac{\Gamma(\sum_j (n_{ij}+\alpha_{ij}))}{\prod_j \Gamma(n_{ij}+\alpha_{ij})} \qquad (22)$$

المعادلة (22) تمثل التوزيع اللاحق لطريقة بيز في تقدير مصفوفة الاحتمالات الانتقالية، وبتشبيه المعادلة (22) بتوزيع Dirichlet distribution بالمعلمات $(n_{ij} + \alpha_{ij})$ وعلية فأن المعدل والتباين للتوزيع اللاحق هو:

$$E[P_{ij}] = \frac{n_{ij}+\alpha_{ij}}{\sum_j (n_{ij}+\alpha_{ij})} \qquad (23)$$

$$Var[P_{ij}] = \frac{(n_{ij}+\alpha_{ij})(\sum_j (n_{ij}+\alpha_{ij}) - (n_{ij}+\alpha_{ij}))}{(\sum_j (n_{ij}+\alpha_{ij}))^2 (\sum_j (n_{ij}+\alpha_{ij})+1)} \qquad (24)$$

ان الافتراض الشائع حسب طريقة بيز لتقدير مصفوفة الاحتمالات الانتقالية يكون بفرض

$$\alpha_{ij} = n_{ij} + 1 \qquad (25)$$

$$\alpha_{i.} = n_{i.} + k \qquad (26)$$

وبتعويض عن المعادلة (25) والمعادلة (26) في معادلة (23) والمعادلة (24) نحصل على المعدل والتباين لطريقة بيز بالشكل( Strelioff, Crutchfield, & Hübler, 2007):

$$E[P_{ij}] = \frac{2n_{ij}+1}{\sum_j (2n_{ij}+1)} \qquad (27)$$

$$Var[P_{ij}] = \frac{(2n_{ij}+1)[(2n_{i.}+k) - (2n_{ij}+1)]}{(2n_{i.}+k)^2\, (2n_{i.}+k+1)} \qquad (28)$$

عندما تكون قيمة $\alpha_{ij}$ تقترب من قيمة $n_{ij}$ ، علية فأن تباين طريقة الإمكان الأعظم MLE ويكون بالشكل((Singer et al., 2014 ):

$$Var[P_{ij}] = \frac{2n_{ij}\,(2n_{i.} - 2n_{ij})}{(2n_{i.})^2\, (2n_{i.}+1)} \qquad (29)$$

وعلية فأن مقدر طريقة بيز للمصفوفة الانتقالية لسلسلة ماركوف هو:

$$P_{ij}^{(Bayes)} = \frac{2n_{ij}+1}{\sum_j (2n_{ij}+1)} \qquad (30)$$

**3.3 طريقة مقترحة لتقدير الاحتمالات الانتقالية The Proposed Method for Estimating the Transition Probabilities :**

إن الأسلوب المقترح لتقدير الاحتمالات الانتقالية هو إجراء تعديل على طريقة بيز في تقدير الاحتمالات الانتقالية، وذلك للوصول الى احتمالات انتقالية بأقل تباين، وبالاعتماد على أسلوب بيز تفترض هذه الطريقة بان قيم $\alpha_{ij}$ في الاحتمال الأولي (والذي يتبع توزيع Dirichlet distribution) يتم تقديرها من خلال توزيع Dirichlet distribution ، لذا يكون الاحتمال الاولي على النحو الآتي((Singer et al., 2014):

$$P(\theta) = \prod_i \frac{\Gamma(\sum_j \alpha_{ij})}{\prod_j \Gamma(\alpha_{ij})} \prod_j P_{ij}^{\alpha_{ij}-1} \qquad (31)$$

ويكون الاحتمال الشرطي للبيانات بالنظر لمعلمات الأنموذج على النحو الآتي:

$$P(D|\theta) = P(x_1) \prod_i \prod_j P_{ij}^{n_{ij}} \qquad (32)$$

واحتمال بيانات الأنموذج يكون على النحو الآتي:

$$P(D) = \int P(D|\theta)\, P(\theta)\, d\theta \qquad (33)$$

وبتعويض المعادلة (31) والمعادلة (32) في المعادلة (33) يتم الحصول على:

$$P(D) = P(x_1)\prod_i \frac{\Gamma(\sum_j \alpha_{ij})}{\prod_j \Gamma(\alpha_{ij})} \int \prod_j P_{ij}^{n_{ij}+\alpha_{ij}-1}\, d\theta$$

$$\int \prod_j P_{ij}^{n_{ij}+\alpha_{ij}-1}\, d\theta = \frac{\prod_j \Gamma(n_{ij}+\alpha_{ij})}{\Gamma(\sum_j (n_{ij}+\alpha_{ij}))}$$

$$P(D) = P(x_1)\prod_i \frac{\Gamma(\sum_j \alpha_{ij})\prod_j \Gamma(n_{ij}+\alpha_{ij})}{\prod_j \Gamma(\alpha_{ij})\Gamma(\sum_j (n_{ij}+\alpha_{ij}))} \quad (34)$$

يتم الحصول على التوزيع اللاحق للطريقة المقترحة في تقدير مصفوفة الاحتمالات الانتقالية من خلال تعويض المعادلة (31) والمعادلة (32) والمعادلة (34) في المعادلة (14):

$$P(\theta|D) = \prod_i \prod_j P_{ij}^{n_{ij}+\alpha_{ij}-1} \frac{\Gamma(\sum_j (n_{ij}+\alpha_{ij}))}{\prod_j \Gamma(n_{ij}+\alpha_{ij})} \quad (35)$$

إن المعادلة (35) تمثل التوزيع اللاحق للطريقة المقترحة في تقدير الاحتمالات الانتقالية، وبتشبيه المعادلة (35) بتوزيع Dirichlet distribution بالمعلمات $(n_{ij} + \alpha_{ij})$ وعلية فأن المعدل والتباين للتوزيع اللاحق هما:

$$E[P_{ij}] = \frac{n_{ij}+\alpha_{ij}}{\sum_j (n_{ij}+\alpha_{ij})} \quad (36)$$

$$Var[P_{ij}] = \frac{(n_{ij}+\alpha_{ij})(\sum_j (n_{ij}+\alpha_{ij}) - (n_{ij}+\alpha_{ij}))}{(\sum_j (n_{ij}+\alpha_{ij}))^2 (\sum_j (n_{ij}+\alpha_{ij})+1)} \quad (37)$$

وعلية فأن مقدر الطريقة المقترحة للاحتمالات الانتقالية لسلسلة ماركوف هو:

$$P_{ij}^{(\hat{\alpha})} = \frac{n_{ij}+\alpha_{ij}}{\sum_j (n_{ij}+\alpha_{ij})} \quad (38)$$

وتم اقتراح طريقة ذكائية لتقدير المعلمة $\alpha_{ij}$ متمثلة بخوارزمية سرب الجسيمات (PSO) وكما يأتي:

### 1.3.3 استخدام خوارزمية سرب الجسيمات لتقدير $\alpha_{ij}$

تعد خوارزمية سرب الجسيمات Particle Swarm Optimization (PSO) من الطرائق الذكائية، لما لها من فائدة كبيرة، وتم في الآونة الأخيرة استخدام هذه الخوارزمية في التطبيقات العملية من اجل الحصول على مقدر بدقة عالية وباقل جهد ووقت((Soyer & Tarimcilar, 2008).

تم اقتراح خوارزمية (PSO) في عام 1995 من قبل الباحثين James Kennedy وRussel Eberhart لحل مشكلة التحسين المستمرة والغير مقيدة.

تعد خوارزمية سرب الجسيمات (PSO) من الطرق الذكائية، ولما لها فائدة كبيرة تم في الآونة الأخيرة استخدام هذه الخوارزمية في التطبيقات العملية من اجل الحصول على مقدر بدقة عالية وباقل جهد ووقت

تستلهم خوارزمية سرب الجسيمات من الأمثلة البيولوجية للسلوك الطبيعي والجماعي لمجتمع الحيوانات والحشرات والمخلوقات التي تعيش في مجموعات، ومن هذه الكائنات الدبابير والنحل والنمل الأبيض والاوز ومن مجتمع الحيوانات مجاميع الأسماك واسراب الطيور. يتم استخدام ذكاء السرب لوصف الأنظمة لتحقيق الحالة المثلى، حيث يتم اتخاذ القرارات في السرب بشكل لامركزي من قبل الافراد على أساس المعلومات التي يتم الحصول عليها من البيئة المحيطة بهم((Fister, Fister Jr, Yang, & Brest, 2013).

تهدف خوارزمية PSO إلى إيجاد الحل الأمثل من خلال التحديث المتكرر لموضع وسرعة كل جسيم بناءً على حركة الجسيمات في السرب

ويكون أساس عمل هذه الخوارزمية هي الجسيمات حيث تقوم بمحاكاة السلوك الطبيعي لأسراب الجسيمات في برنامج حاسوبي، حيث يتم في البداية تهيئة الجسيمات ويكون كل جسيم من هذه الجسيمات له سرعة وموضع خاص به، وتطير تلك الجسيمات في فضاء البحث ويتم تعديل السرعة الخاصة بالجسيمات عن طريق التحكم في موقعها الحالي ، وسرعتها ويتم تحديثها في كل تكرار للخوارزمية لذا تكون تلك الجسيمات لها ميل لتطير نحو الحل الأفضل في فضاء البحث وتكون معادلات التحديث لموضع وسرعة الجسيم هي كما يلي:

$$V_i^{t+1} = \omega V_i^t + c_1 r_1 (P_{best} - X_i^t) + c_2 r_2 (g_{best} - X_i^t) \quad (39)$$

$$X_i^{t+1} = X_i^t + V_i^{t+1} \quad (40)$$

إذ إن:

$$X_i^0 \sim U(X_{Min}, X_{Max})$$

وعلية فإن:

$$X_i^0 = X_{Min} + r_i\,(X_{Max} - X_{Min}) \;\;,\;\; r_i \sim U(0\,,1)$$

إذ إن

$V_i^{t+1}$ : هي سرعة الجسيم عند التكرار $k$

$X_i^{t+1}$ موضع الجسيم عند التكرار $k$

$c_1$ , $c_2$ معاملات تسريع تنظم إلى أي مدى يمكن للجسيم أن يتحرك في تكرار واحد

$r_1$ , $r_2$ : ارقام عشوائية من التوزيع المنتظم

$\omega$ : ثابت موجب يمثل وزن القصور الذاتي

لذا يمكن تلخيص الخطوات الرئيسة لخوارزمية PSO على النحو الآتي:

1. تهيئة الموضع عن طريق تعيين موضع عشوائي لكل جسيم.
2. حساب قيمة ملائمة لكل جسيم ضمن السرب.
3. يتم تحديث أفضل وضع محلي إذا كان أفضل من السابق.
4. يتم تحديث أفضل مركز عالمي إذا كان أفضل من السابق.
5. حساب سرعة كل جسيم باستخدام المعادلة (39).
6. تحديث موضع الجسيمات باستخدام المعادلة (40).
7. يتم تكرار الخطوات (2–6) حتى يتم تحقق شرط الانتهاء.

**4. الجانب التطبيقي:**

للمعلوماتية الحيوية استخدام واسع النطاق في دراسة الجينوم للكائنات الحية في تحديد سلسلة الحامض النووي الرايبوزي منقوص الاوكسجين DNA. وان علم المعلوماتية الحيوية Bioinformatics يعتمد على كل من علم الإحصاء والرياضيات والحاسوب والكيمياء والطب في تحليل البيانات. واستخدم علم المعلوماتية الحيوية Bioinformatics في تحديد متسلسلات الحامض النووي الرايبوزي منقوص الاوكسجين DNA، والحامض النووي هو مركب كيميائي معقد التركيب مسؤول عن تحديد الصفات الوراثية وموجود في جميع الكائنات الحية. يوجد الحامض النووي الرايبوزي منقوص الاوكسجين DNA في نواة الخلية ضمن الكروموسومات التي تتكون من الشبكة الكروماتينية، وعرف الحمض النووي الرايبوزي منقوص الاوكسجين في عام 1900 بأنه شريط طويل متكون من أربع قواعد نيتروجينية والتي تكون على نوعين هي:

1. قواعد بيورين Purines وهي :
   - ادنين Adenine ويرمز له بالرمز A.
   - كوانين Guanine ويرمز له بالرمز G.
2. قواعد بيرميدين Pyrimidines وهي :
   - سايتوسين Cytosine ويرمز له بالرمز C.
   - الثايمين Thiamin ويرمز له بالرمز T.

ويتكون شريط الـ DNA من ارتباط القواعد النيتروجينية فيما بينها اذ يرتبط الثايمين مع الادنين، ويرتبط السايتوسين مع الكوانين. تم الحصول على بيانات بكتيريا القولون الاشريكية Escherichia Coli (E.Coli) من الموقع الالكتروني الخاص بالمركز الوطني لمعلومات التكنولوجيا الحيوية من خلال الرابط https://www.ncbi.nlm.nih.gov ، حيث يوفر هذا الموقع قاعدة من البيانات والتي تكون متاحة للباحثين لأغراض التطوير والبحث العلمي، وتم اختيار سلسلة الجين الخاصة بـ E.Coli بطول 1039 قاعدة نيتروجينية

كجانب تطبيقي للدراسة وذلك لاهميتها في الأبحاث الطبية ولغرض اكتشاف وتصنيع العلاجات من خلال معرفة الشكل النهائي لسلسلة الجين الخاص بها.

وباستخدام البرنامج المعد لهذا الغرض باللغة البرمجية MATLAB R2021a تم تكوين مصفوفة التكرارات لسلسلة جين E.Coli ، والتي تضم اعداد الانتقالات بين القواعد النيتروجينية الأربع.

$$F = \begin{array}{c} \\ A \\ T \\ C \\ G \\ f_{.j} \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{bmatrix} 86 & 65 & 68 & 46 \\ 47 & 76 & 42 & 81 \\ 74 & 49 & 65 & 73 \\ 57 & 56 & 86 & 68 \end{bmatrix} \\ 264 \quad 246 \quad 261 \quad 268 \end{array} \begin{array}{c} f_{i.} \\ 265 \\ 246 \\ 261 \\ 267 \\ N = 1039 \end{array}$$

وتم اختبار سلسلة جين E.Coli وبفرض كل من فرضية العدم والفرضية البديلة والتي تنص:

$H_0$: سلسلة جين E.Coli لا تمثل سلسلة ماركوف

$H_1$: سلسلة جين E.Coli تمثل سلسلة ماركوف

وان المختبر الاحصائي الذي يختبر هل ان سلسلة جين E.Coli تمثل سلسلة ماركوف ام لا يتبع توزيع $\chi_n$ بدرجة حرية $(s-1)^2$

حيث

$s$ : تمثل عدد الحالات في مصفوفة التكرارات

$n_{ij}$ : تمثل المشاهدة $i,j$ في مصفوفة التكرارات

$n_{i.}$ : تمثل مجموع الصف في مصفوفة التكرارات

$N$ : تمثل مجموع مصفوفة التكرارات

$n_{.j}$ : تمثل مجموع العمود في مصفوفة التكرارات

وبمقارنة قيمة $\chi_n$ المحسوبة والتي تساوي (1361.4) مع قيمة $\chi_n$ الجدولية والتي تساوي (27.88)، نرفض فرضية العدم ونقبل الفرضية البديلة، أي ان سلسلة جين E.Coli تمثل سلسلة ماركوف.

**5. مناقشة النتائج**

تم تقدير مصفوفة الاحتمالات الانتقالية لسلسلة الجين E.Coli باستخدام طريقة الإمكان الأعظم وطريقة بيز والطريقة المقترحة، الجدول التالي يبين احتمالات الانتقال بكل من طريقة الإمكان الأعظم وطريقة بيز والطريقة المقترحة.

جدول (1) : الاحتمالات الانتقالية لسلسلة جين E.Coli باستخدام طرائق تقدير مختلفة

| Transition | probability of Transition Method | | |
|---|---|---|---|
| | MLE | Proposed Method | Bayes |
| AA | 0.3245 | 0.3225 | 0.324 |
| AT | 0.2453 | 0.2460 | 0.324 |
| AC | 0.2566 | 0.2573 | 0.2566 |
| AG | 0.1736 | 0.1743 | 0.1742 |
| TA | 0.1911 | 0.1918 | 0.1915 |
| TT | 0.3089 | 0.3109 | 0.3085 |
| TC | 0.1707 | 0.1715 | 0.1714 |
| TG | 0.3293 | 0.3258 | 0.3286 |
| CA | 0.2835 | 0.2818 | 0.2833 |
| CT | 0.1877 | 0.1884 | 0.1882 |
| CC | 0.249 | 0.2496 | 0.249 |
| CG | 0.2797 | 0.2802 | 0.2795 |
| GA | 0.2135 | 0.2141 | 0.2138 |
| GT | 0.2097 | 0.2105 | 0.21 |

| | | | |
|---|---|---|---|
| GC | 0.3221 | 0.3201 | 0.3216 |
| GG | 0.2547 | 0.2554 | 0.2546 |

والجدول التالي يبين تباين احتمالات الانتقال بكل من طريقة الإمكان الأعظم وطريقة بيز والطريقة المقترحة.

جدول (2) : تباين احتمالات الانتقال لسلسلة جين E.Coli باستخدام طرائق تقدير مختلفة

| Transition | Variance of Transition | | | Minimum of Variance |
|---|---|---|---|---|
| | Method | | | |
| | MLE | Proposed Method | Bayes | |
| AA | 0.00041282 | 0.0003907 | 0.00040937 | Proposed Method |
| AT | 0.00034862 | 0.00033167 | 0.00034605 | Proposed Method |
| AC | 0.00035924 | 0.00034169 | 0.00035651 | Proposed Method |
| AG | 0.00027016 | 0.00025738 | 0.00026883 | Proposed Method |
| TA | 0.0003135 | 0.00028806 | 0.00031156 | Proposed Method |
| TT | 0.00043306 | 0.00039811 | 0.00042921 | Proposed Method |
| TC | 0.00028719 | 0.00026403 | 0.00028572 | Proposed Method |
| TG | 0.00044797 | 0.00040816 | 0.00044393 | Proposed Method |
| CA | 0.00038841 | 0.00033898 | 0.00038525 | Proposed Method |
| CT | 0.00029157 | 0.00025607 | 0.00028992 | Proposed Method |
| CC | 0.00035759 | 0.0003137 | 0.00035488 | Proposed Method |
| CG | 0.00038521 | 0.00033775 | 0.0003821 | Proposed Method |
| GA | 0.00031385 | 0.00030031 | 0.00031181 | Proposed Method |
| GT | 0.00030981 | 0.00029664 | 0.00030783 | Proposed Method |
| GC | 0.00040813 | 0.0003885 | 0.00040475 | Proposed Method |
| GG | 0.0003548 | 0.00033944 | 0.00035214 | Proposed Method |

**6. الاستنتاجات**

من مناقشة النتائج تم التوصل الى

1- الطريقة المقترحة افضل مقارنةً مع طريقة الإمكان الاعظم وطريقة بيز في تقدير المصفوفة الانتقالية بالاعتماد على التباين.

2- من مصفوفة الاحتمالات الانتقالية المقدرة حسب الطريقة المقترحة لسلسلة جين E.Coli تبين ان القاعدة النيتروجينية T تظهر باحتمالية اكبر بعد ظهور القاعدة النيتروجينية G يليها القواعد النيتروجينية AA ثم القواعد النيتروجينية GC ثم القواعد النيتروجينية TT.

3- من مصفوفة الاحتمالات الانتقالية المقدرة حسب طريقة الإمكان الاعظم لسلسلة جين E.Coli تبين ان القاعدة النيتروجينية G تظهر باحتمالية اكبر بعد ظهور القاعدة النيتروجينية T يليها القواعد النيتروجينية AA ثم القواعد النيتروجينية GC ثم القواعد النيتروجينية TT.

4- من مصفوفة الاحتمالات الانتقالية المقدرة حسب طريقة بيز لسلسلة جين E.Coli تبين ان القاعدة النيتروجينية G تظهر باحتمالية اكبر بعد ظهور القاعدة النيتروجينية T يليها كل من القواعد النيتروجينية AA و AT يليها القواعد النيتروجينية GC ثم القواعد النيتروجينية TT.

**7. توصيات**

1- نوصي باستخدام طرائق ذكائية لتقدير مصفوفة الاحتمالات الانتقالية ومقارنة النتائج مع الطرق التقليدية.

2- نوصي الجهة المستفيدة وهم الباحثين في علوم الحياة باعتماد طريقة بيز لتقدير المصفوفة الانتقالية لسلسلة جين E.Coli وذلك لاهميتها في الأبحاث الطبية ولغرض اكتشاف وتصنيع العلاجات من خلال معرفة الشكل النهائي لسلسلة الجين.

**Reference**


1. Demirel, A., & Çelik, H. (2018). DIRICHLET DISTRIBUTION AND ESTIMATION OF PARAMETERS. *ADVANCES APPLICATIONS IN STATISTICS, 53*(4).
2. Fister, I., Fister Jr, I., Yang, X.-S., & Brest, J. (2013). A comprehensive review of firefly algorithms. *%J Swarm Evolutionary Computation* .46-34 ,13
3. Guo, H. R., Liao, H., Zhao, W., & Mettas, A. J. I. T. o. R. (2007). A new stochastic model for systems under general repairs. *56*(1), 40-49.
4. Kalbfleisch, J. D., & Lawless, J. F. (1984). Least-squares estimation of transition probabilities from aggregate data. *Canadian Journal of Statistics, 12*(3), 169-182.
5. Lee, T. C., Judge, G., & Zellner, A. (1968). Maximum likelihood and Bayesian estimation of transition probabilities. *Journal of the American Statistical Association, 63*(324) .1179-1162 ,
6. Lin, J. (2016). On the dirichlet distribution. *Department of Mathematics Statistics, Queens University* .11-10
7. Ma, J., Yu, X., Symanski, E., Doody, R., & Chan, W. (2014). A Bayesian Approach in Estimating Transition Probabilities of a Discrete-time Markov Chain for Ignorable Intermittent Missing Data. *Communications in Statistics-Simulation and Computation, 45*(7), 2598-2616.
8. Singer, P., Helic, D., Taraghi, B., & Strohmaier, M. J. P. o. (2014). Detecting memory and structure in human navigation patterns using Markov chain models of varying order. *9*(7), e102070.
9. Soyer, R., & Tarimcilar, M. M. (2008). Modeling and analysis of call center arrival data: A Bayesian approach. *%J Management Science, 54*(2), 266-278.
10. Strelioff, C. C., Crutchfield, J. P., & Hübler, A. W. J. P. R. E. (2007). Inferring Markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling. *76*(1), 011106.
11. Wang, G. (2010). ML estimation of transition probabilities in jump Markov systems via convex optimization. *IEEE Transactions on Aerospace Electronic Systems* .1502-1492 ,(3)46
12. Yang, J., & Sha, Q. (2011). *Research and application by Markov chain operators in the mobile phone market.* Paper presented at the 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC).

# Estimating the Transitional Probabilities of the E.Coli Gene Chain by Maximum Likelihood Method and Bayes Method

Alaa Farhan Ahmed [1] , Muthanna Subhi Sulaiman[2]
[1,2]Department of Informatics & Statistic, College of Computer Science and Mathematics , University of Mosul ,Mosul ,Iraq

**Abstract**

The transition matrix estimators of the Markov chain are not accurate and the transition matrix is considered given. There are many methods that are used to estimate the transition probabilities matrix for different cases, the most famous of which is the Maximum Likelihood Method, In order to find a good estimator for the transition probabilities matrix of the Markov chain, a Bayes method and a Proposed Method was used in this paper, to reach the transition probabilities with the least variance, The Escherichia Coli (E.Coli) gene chain was chosen as an applied aspect of the study due to its importance in medical research and for the purpose of discovering and manufacturing treatments by knowing the final form of its gene chain. After testing the E.Coli gene chain, it was found that is represents a Markov chain, and then both the transition probabilities matrix and the transition probabilities variance were estimated used Proposed Method and Bayes method and Maximum Likelihood Method, and it was found that the Proposed Method for transitional probabilities is better than the Bayes method and Maximum Likelihood Method dependence on the variance.

**Keywords:** Transitional Probabilities, Markov Chain, MLE, Bayes, DNA