المجلة العراقية للعلوم الإحصائية

www.stats.mosuljournals.com

توظيف نموذج ماركوف المخفي في تحديد نوعية القاعدة النيتروجينية المستبدلة لسلسلة الجين MT-ND5 للإنسان والفئران

سرى محمد جمال الدين حسين و مثنى صبحي سليمان

قسم الاحصاء والمعلوماتية ، كلية علوم الحاسوب والرياضيات، جامعة الموصل ، الموصل ، العراق

معلومات النشر

تاريخ المقالة:
تم استلامه في 1 كانون الثاني 2023
تم القبول في 12 اذار 2023
متاح على الإنترنت في 1 كانون الاول 2023

الكلمات الدالة:
نموذج ماركوف المخفي، خوارزمية فيتربي، سلسلة الجين MT-ND5 للإنسان والفئران.

المراسلة:
سرى محمد جمال الدين
sura.alalwlia@ uomosul.edu.iq

الخلاصة

تم تطوير نماذج ماركوف المخفية لتحليل بيانات المعلوماتية الحيوية التي استقطبت اهتمامات الباحثين لأهميتها البالغة في حياة الكائنات الحية، كان هدف البحث تحديد نوعية القاعدة النيتروجينية المستبدلة لسلسلة الجين MT-ND5 للإنسان والفئران، اثبتت الخوارزمية المقترحة في استخدام خوارزمية Viterbi في نموذج ماركوف المخفي انها جيدة في تحديد نوعية القاعدة النتروجينية المستبدلة لسلسلة الجينMT-ND5 الخاصة بالإنسان والفئران وذلك بالاعتماد على النسب العالية للتطابق التي تم الحصول عليها وعلى مجموع مربعات الخطأ المنخفضة. وتم تصميم برنامج حاسوبي لهذا الغرض وتمت برمجة الخوارزمية بلغة MATLAB R2017b، ومن التطبيق العملي للخوارزمية يتبين أن نموذج ماركوف المخفي هو نهج قوي بشكل خاص لتحديد نسبة التطابق التي تصل إلى دقة تصنيف عالية.

DOI: https://doi.org/10.33899/iqjoss.2023.0178691 , ©Authors, 2023, College of Computer and Mathematical Science, University of Mosul.

1. المقدمة Introduction:

شهد القرن العشرين تقدماً هائلاً في الأساليب العلمية المستخدمة في البحث العلمي في ميادين المعرفة كافة، وأصبح الاهتمام ملحوظاً بشكل أوسع في دراسة الانظمة التي تتغير مع الزمن بشكل عشوائي. ويطلق على النماذج الرياضية لمثل هذه الانظمة بالعمليات التصادفية والتي تضم مجموعة كبيرة من النماذج منها نموذج ماركوف المخفي Hidden Markov Model (HMM)، الذي يعد من النماذج التصادفية المهمة والذي تم تطبيقه في البدء كنموذج احصائي لتمييز الكلام Speech Recognition والكتابة اليدوية Handwriting، بسبب قدرته الكبيرة على التكيف مع المشكلة فضلا عن البراعة في التعامل مع الاشارات المتسلسلة [3].

2. نموذج ماركوف المخفي Hidden Markov Model (HMM)

إن مفهوم نموذج ماركوف المخفي HMM وخوارزمياته مستلهم أساساً من نماذج رياضية معروفة باسم العالم الذي اكتشفها وهو Andrei Markov. وقد ظهرت هذه النماذج في مستهل القرن العشرين وأطلق عليها نماذج ماركوف Markov Models، وهذا يدل على أن نموذج ماركوف المخفي ما هو إلا امتداد لنموذج ماركوف الاعتيادي [1, 7] . ويعد نموذج ماركوف المخفي مجموعة منتهية من الحالات، وكل حالة تقترن بتوزيع احتمالي. وبشكل عام تتولد الحالة الناتجة طبقاً للاحتمالات المقترنة بالحالة حيث توجد احتمالات

ناجحة فقط ولا توجد حالة ظاهرة يمكن مشاهدتها، لذا تكون الحالات مخفية، أي ان نموذج ماركوف المخفي أداة احصائية قوية تستخدم للتنبؤ بسلسلة الحالة من خلال سلسلة المشاهدات. وتعد معلمة نموذج ماركوف المخفي $\lambda = (A, B, \pi)$ امتداد لمعلمة نموذج ماركوف الاعتيادي $\lambda = (A, \pi)$. وقد بدأ استخدام نموذج ماركوف المخفي في النصف الثاني من ثمانينيات القرن العشرين بتحليل المتتابعات الحيوية Biological Sequences، وبخاصة متتابعات الـ DNA. ومنذ ذلك الحين فرض نموذج ماركوف المخفي وجوده في مجال المعلوماتية الحيوية Bioinformatics الذي يهتم بقواعد البيانات الحيوية والوراثية وادارتها وتطويرها [10]. والعناصر المهمة لنموذج ماركوف المخفي هي: [6]

1) سلسلة المشاهدات (O):

$$O = \{o_1, o_2, \dots, o_T\}$$

اذ ان T تمثل طول سلسلة المشاهدات، ومؤشر رموز المشاهدات هو (V)، إذ ان:

$$V = 1,2, \dots, M$$

و M هو عدد رموز المشاهدات

$$o_i \in V\ ;\ i = 1,2, \dots, T$$

2) سلسلة الحالات المخفية (Q):

$$Q = \{q_1, q_2, \dots, q_N\}$$

اذ ان N تمثل عدد الحالات المخفية في النموذج والتي تكافئ فضاء الحالة (S) في نموذج ماركوف وكما يأتي:

$$S = \{s_1, s_2, \dots, s_N\}$$

3) مصفوفة الاحتمالات الانتقالية Transition Probability Matrix (A): وتمثل عناصرها التوزيع الشرطي للحالة الانتقالية، اذ ان:

$$A = \{a_{ij}\}\ ;\ i, j = 1,2, \dots, N$$
$$a_{ij} = P\{q_{t+1} = S_j | q_t = S_i\}$$

اذ ان a_{ij} تمثل عناصر المصفوفة A وتحقق الشروط الاتية:

1. $a_{ij} \geq 0$
2. $\sum_{j=1}^{N} a_{ij} = 1$

4) مصفوفة الإصدارات Emission Matrix (B): وتمثل مصفوفة احتمالية رابطة بين الحالات المخفية والمشاهدات.

$$B = \{b_j(k)\}\ ;\ j = 1,2, \dots, N, k = 1,2, \dots, M$$
$$b_j(k) = P\{o_t = v_k | q_t = S_j\}$$

إذ ان v_k يمثل رمز المشاهدة k ، وتحقق الشروط الاتية:

1. $b_j \geq 0$
2. $\sum_{k=1}^{M} b_j(k) = 1$

5) متجه الحالة الابتدائية The Initial State (π): وتمثل الحالات الابتدائية لنموذج ماركوف المخفي، إذ ان:

$$\pi = \{\pi_i\}\ ;\ i = 1,2, \dots, N$$

$$\pi_i = P_r\{q_1 = S_i\}$$

إذ ان π_i تمثل عناصر المتجه π وتحقق الشروط الاتية:

1. $\pi_i \geq 0$
2. $\sum_{j=1}^{N} \pi_j = 1$

3. المسائل الاساسية لنموذج ماركوف المخفي The Basic Problems for HMM

هناك ثلاث مسائل أساسية عند دراسة نموذج ماركوف المخفي:

- **مسالة التقييم Evaluation Problem**

تعمل مسألة التقييم على حساب احتمالية سلسلة المشاهدات $P(O|\lambda)$ للنموذج عندما يكون النموذج $\lambda = (A, B, \pi)$ هو المعطى. أي يتم دراسة إمكانية احتمالية سلسلة المشاهدة بشكل كفوء عندما يكون النموذج معطى، وتحل عن طريق الخوارزمية الأمامية – الخلفية Forward- Backward Algorithm [5].

- **مسألة الشفرة Decoding Problem**

تعمل مسألة الشفرة على إيجاد سلسلة الحالة المثلى$Q = \{q_1, q_2, \dots, q_T$ عندما تكون سلسلة المشاهدات (O) والنموذج $\lambda = (A, B, \pi)$ معطى. وتحل هذه المسألة عن طريق خوارزمية فيتربي Viterbi Algorithm [2].

- **مسألة التدريب Training Problem**

تعمل مسألة التدريب على إعادة تقدير معلمات النموذج $\lambda = (A, B, \pi)$ التي تعظم من إمكانية $P(O|\lambda)$ عندما تكون سلسلة المشاهدة $O = \{o_1, o_2, \dots, o_T\}$ معطى. وتحل هذه المسألة عن طريق خوارزمية بوم ولتش Baum-Welch Algorithm [8].

4. حل مسالة الشفرة باستخدام خوارزمية فيتربي Viterbi

خوارزمية Viterbi هي خوارزمية تعمل على إيجاد أفضل سلسلة حالة بشكل وحيد، والمتغيرات الأساسية لهذه الخوارزمية هي :[4,9]

- المتغير $S_t(i)$: يمثل أعلى احتمالية على طول المسار الوحيد في الحالة (i) عند الزمن (t) والذي يساوي احتمالية سلسلة الحالة الجزئية الأكثر احتمالاً بالنسبة لسلسلة المشاهدات المنتهية في الحالة (i) ويمكن التعبير عنه رياضيا وكما يأتي:

$$S_t(i) = Max_{q1,q2,..,qt-1} P\{q_1, q_2, \dots, q_{t=}S_i, O_1, O_2, \dots, O_{t|\lambda}\} \quad (1)$$

إذ ان

$$t = 1,2, \dots, T \; ; \; i = 1,2, \dots, N$$

- المتغير $\psi_t(i)$: يعمل هذا المتغير على حفظ تتابع الأثر Keep Track للمسار الفعلي.

إن خطوات سير خوارزمية Viterbi يمكن أن تمثل بالشكل الآتي [2]:

1) البداية Initialization

$$S_1(i) = \pi_i \; b_i(O_1) \; ; \; i = 1,2, \dots, N \quad (2)$$

$$\psi_1(i) = 0 \quad (3)$$

2) الحث (التعاقب) Induction

$$S_t(j) = Max_{1 \le i \le N} \{ S_{t-1}(i) a_{ij} \} b_j(O_t) \; ; i = 1,2, \dots, N \; ; \; t = 1,2, \dots, T \quad (4)$$

$$\psi_t(j) = \arg Max_{1 \le i \le N} \{ S_{t-1}(i) a_{ij} \} \; ; \; j = 1,2, \dots, N \; ; \; t = 1,2, \dots, T \quad (5)$$

$\arg Max$:الوسيط الاعظمي يعرف في الرياضيات على انه وسيط (دخل الدالة) التي نعطي اكبر قيمة (حدود عليا وحدود دنيا (الحد الأدنى) للدالة في الخرج.

3) النهاية Termination

$$p^*_{\square} = Max_{1 \le i \le N} \{ S_T(i) \} \quad (6)$$

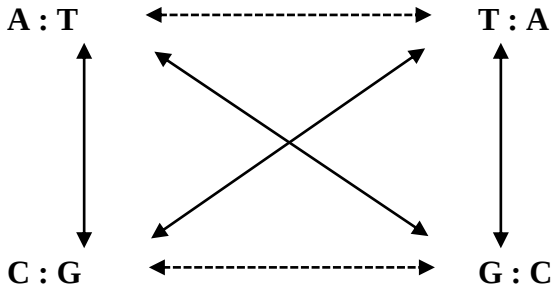
$$q^*_{T\,\square} = \arg Max_{1 \le i \le N} \{ S_T(i) \} \quad (7)$$

4) تراجع سلسلة الحالة المثالية Optimal State Sequence Backtracking

$$q^*_{T\,\square} = \Psi t + 1\{q * t + 1\} \quad ; T = t - 1, t - 2 \; \dots, 2, 1 \quad (8)$$

5. الجانب التطبيقي: تحديد نوعية القاعدة النتروجينة المستبدلة لسلسلة الجين MT-ND5

يمكن تعريف الجين MT-ND5 بانه جين لترميز الجينوم الميتوكوندري للبروتين الخامس NADH-Ubiquinone Oxidoreductase Chain 5، اذا ان البروتين ND5 هو وحدة فرعية لـ NADH dehydrogenase، والذي يقع في الغشاء الداخلي للميتوكوندريا ويمثل اكبر المجمعات الخمسة في سلسلة نقل الالكترون. والشكل الآتي يوضح الاستبدال بين القواعد النتروجينية الأربعة (A, T, C, G) حيث ان اضلاع المربع تمثل طفرات التحول وأقطاره تمثل طفرات الانتقال، أي ان هنالك 12 نوعا من الطفرات



الشكل (1): الاستبدال بين القواعد النيتروجينية حيث ان اضلاع المربع تمثل طفرات التحول والاقطار تمثل طفرات الانتقال

.

وقد تم التطبيق على طفرات الاستبدال على سلسلة الجين MT-ND5 الخاصة بالإنسان والفئران وذلك للمقارنة بين نسبة التطابق للسلسلتين والتي يمكن الحصول عليهما من عملية الاستبدال، وقد تم استخدام خوارزمية Viterbi لتحديد نوعية القاعدة النيتروجينية المستبدلة لسلسلة الجين MT-ND5. كما تم اقتراح خوارزمية لتحديد نوعية القاعدة النتروجينية المستبدلة لسلسلة الجين MT-ND5 لكل من الإنسان والفئران وكما يأتي:

الخوارزمية المقترحة The Suggested Algorithm

الخطوة (1): ترميز القواعد النيتروجينية الأربعة من خلال تحويل الرموز الحرفية إلى أرقام والتي تشكل سلسلة الحامض النووي الرايبي منقوص الأوكسجين وكما يأتي:

$$A = 1, T = 2, C = 3, G = 4$$

الخطوة (2): تعريف عناصر نموذج ماركوف المخفي $\lambda = (A, B, \pi)$ ، اذ ان π تمثل متجه الحالة الابتدائية والذي أبعاده $1 * N$ ، وان $N = 4$ يمثل عدد الحالات. أما A فتمثل مصفوفة الاحتمالات الانتقالية بين الحالات المخفية والتي تكون أبعادها بشكل عام $(N * N)$. وB تمثل مصفوفة احتمالية رابطة بين الحالات المخفية والمشاهدات (مصفوفة الإصدارات) والتي أبعادها $(N * M)$، إذ أن $M = 3$.

الخطوة (3): تتضمن هذه الخطوة 12 مرحلة، وعلى النحو الآتي:

المرحلة الأولى: استبدال القاعدة النيتروجينية A ووضع بدل هذه القاعدة المستبدلة الرمز T.

المرحلة الثانية: استبدال القاعدة النيتروجينية A ووضع بدل هذه القاعدة المستبدلة الرمزC.

المرحلة الثالثة: استبدال القاعدة النيتروجينية A ووضع بدل هذه القاعدة المستبدلة الرمز G.

المرحلة االرابعة: استبدال القاعدة النيتروجينية T ووضع بدل هذه القاعدة المستبدلة الرمز C.

المرحلة الخامسة: استبدال القاعدة النيتروجينية T ووضع بدل هذه القاعدة المستبدلة االرمز G.

المرحلة السادسة: استبدال القاعدة النيتروجينية T ووضع بدل هذه القاعدة المستبدلة الرمز A.

المرحلة السابعة: استبدال القاعدة النيتروجينية C ووضع بدل هذه القاعدة المستبدلة الرمز A.

المرحلة الثامنة: استبدال القاعدة النيتروجينية C ووضع بدل هذه القاعدة المستبدلة الرمز T.

المرحلة التاسعة: استبدال القاعدة النيتروجينية C ووضع بدل هذه القاعدة المستبدلة الرمز G.

المرحلة العاشرة: استبدال القاعدة النيتروجينية G ووضع بدل هذه القاعدة المستبدلة الرمز A.

المرحلة الحادية عشر: استبدال القاعدة النيتروجينية G ووضع بدل هذه القاعدة المستبدلة الرمز C.

المرحلة الثانية عشر: استبدال القاعدة النيتروجينية G ووضع بدل هذه القاعدة المستبدلة الرمز T.

الخطوة (4): إيجاد الحالات المخفية المرجحة وذلك باستخدام خوارزمية Viterbi .

الخطوة (5): تقارن سلسلة الحالات الناتجة من الخطوة (4) مع سلسلة الحالات الحقيقة، حيث يتم في هذه الخطوة تقدير نوعية القاعدة النيتروجينية المستبدلة بما يقابلها بسلسلة الحالات الناتجة من الخطوة (4)، ويتم إيجاد متوسط مجموع مربعات خطأ Mean Squares Error (MSE) حسب الصيغة

$$MSE = \frac{1}{L}\sum_{i=1}^{L}(Q - decode)^2 \quad (9)$$

إذ إن Q: تمثل الحالات المخفية الحقيقية

decode: تمثل الحالات المخفية المشفرة، L : تمثل طول السلسلة.

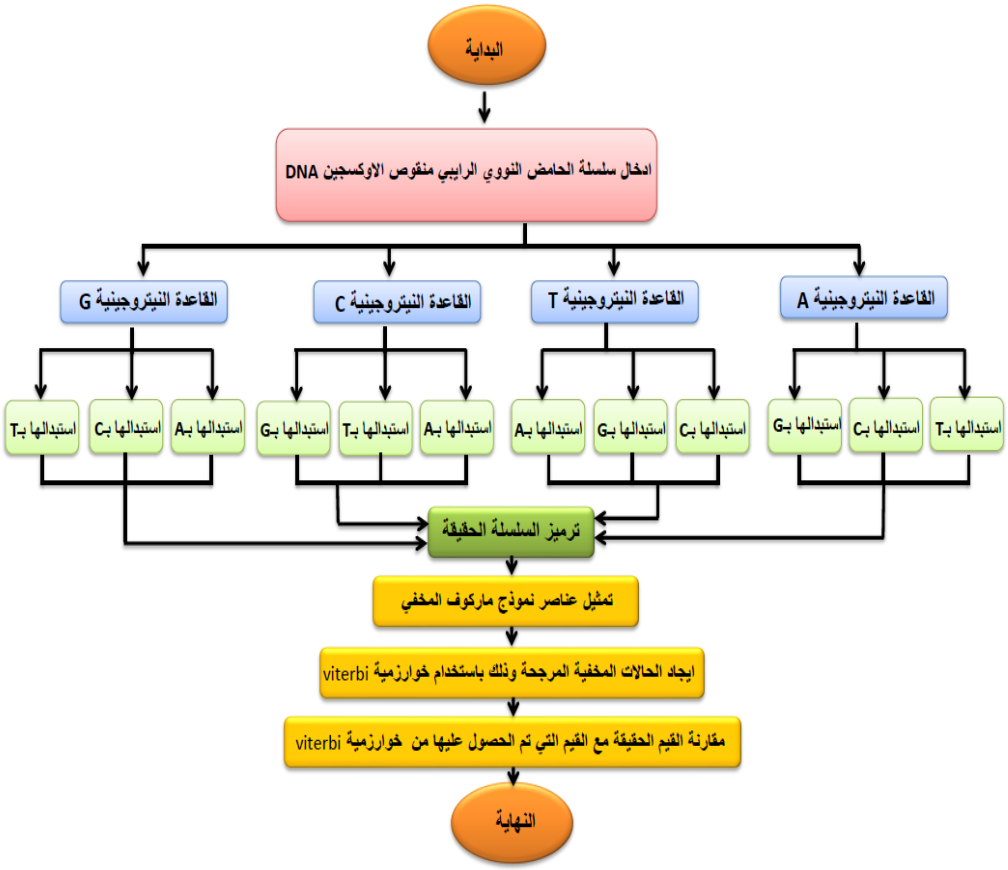
والنسبة المئوية للتطابق Match Ratio (MR) حسب الصيغة

$$MR\% = \left((L - sum(error))/L\right) * 100 \quad (10)$$

إذ إن:

(error): يمثل متجه الأخطاء ذو البعد (L × 1)، ويمثل متجه التعبير المنطقي logical (اما 0 او 1).

والشكل التالي يوضح المخطط الانسيابي للخوارزمية المقترحة لتحديد نوعية القاعدة النتروجينية المستبدلة لسلسلة الجين MT-ND5 الخاصة بالإنسان والفئران:



الشكل (2): المخطط الانسيابي للخوارزمية المقترحة لتحديد نوعية القاعدة النيتروجينية المستبدلة للجين MT-ND5

اولاً: نتائج تطبيق خوارزمية المقترحة على سلسلة الجين MT-ND5 الخاصة بالإنسان

تم اختيار سلسلة الجين MT-ND5 الخاصة بالإنسان من الموقع MT-ND5 mitochondrially encoded NADH dehydrogenase 5 Homo sapiens (human) والتي تتكون من 1812 قاعدة نيتروجينية وذلك لتحديد نوعية القاعدة النتروجينة المستبدلة لسلسلة الجين والتي تم الحصول عليها من موقع NCBI ضمن قاعدة بيانات Data Base في مراكز عالمية متخصصة في الهندسة الوراثية ودراسة عمل الجينات، وباستخدام الخوارزمية المقترحة لتحديد نوعية القاعدة النيتروجينية المستبدلة لسلسة الجين MT-ND5 ، والتي تم برمجتها باستخدام اللغة البرمجية MATLAB R2017b ، وتم استبدال القواعد النتروجينية (A, T, C, G) وكما يأتي:

1) استبدال القاعدة النتروجينية A والتي عددها 551 من سلسلة الجين MT-ND5 بالرمز T، وتم معالجتها باستخدام الخوارزمية المعدة لهذا الغرض، وكانت النتائج كما يأتي:

مصفوفة الاحتمالات الانتقالية (A) هي:

$$A = \begin{bmatrix} 0.3091 & 0.2400 & 0.3018 & 0.1491 \\ 0.3177 & 0.2282 & 0.3647 & 0.0895 \\ 0.3055 & 0.2894 & 0.3408 & 0.0643 \\ 0.2500 & 0.1719 & 0.4219 & 0.1563 \end{bmatrix}$$

مصفوفة الإصدارات (B) هي:

السطر الأول : استبدال A بـ T هو (1) واستبدال A بـ C,G هو (0)

السطر الثاني : استبدال T بـ T هو (1) واستبدال T بـ C,G هو (0)

السطر الثالث: استبدال C بـ C هو (1) واستبدال C بـ T,G هو (0)

السطر الرابع: استبدال G بـ G هو (1) واستبدال G بـ C,T هو (0)

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$MSE = 0.2439$$

$$MR = 75.6071$$

واستبدالها بالرمز C، وتم معالجتها باستخدام الخوارزمية المعدة لهذا الغرض، وكانت النتائج كما يأتي:

مصفوفة الاحتمالات الانتقالية (A) هي:

$$A = \begin{bmatrix} 0.3091 & 0.2400 & 0.3018 & 0.1491 \\ 0.3177 & 0.2282 & 0.3647 & 0.0895 \\ 0.3055 & 0.2894 & 0.3408 & 0.0643 \\ 0.2500 & 0.1719 & 0.4219 & 0.1563 \end{bmatrix}$$

مصفوفة الإصدارات (B) هي:

السطر الأول : استبدال A بـ T هو (0) واستبدال A بـ C هو (1) واستبدال A بـ G هو (0)

السطر الثاني : استبدال T بـ T هو (1) واستبدال T بـ C,G هو (0)

السطر الثالث: استبدال C بـ C هو (1) واستبدال C بـ T,G هو (0)

السطر الرابع: استبدال G بـ G هو (1) واستبدال G بـ C,T هو (0)

$$B = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$MSE = 1.1236$$

$$MR = 71.9095$$

واستبدالها بالرمز G، وتم معالجتها باستخدام الخوارزمية المعدة لهذا الغرض، وكانت النتائج كما يأتي:

مصفوفة الاحتمالات الانتقالية (A) هي:

$$A = \begin{bmatrix} 0.3091 & 0.2400 & 0.3018 & 0.1491 \\ 0.3177 & 0.2282 & 0.3647 & 0.0895 \\ 0.3055 & 0.2894 & 0.3408 & 0.0643 \\ 0.2500 & 0.1719 & 0.4219 & 0.1563 \end{bmatrix}$$

مصفوفة الإصدارات (B) هي:

السطر الأول : استبدال A بـ T هو (0) واستبدال A بـ C هو (0) واستبدال A بـ G هو (1)
السطر الثاني : استبدال T بـ T هو (1) واستبدال T بـ C,G هو (0)
السطر الثالث: استبدال C بـ C هو (1) واستبدال C بـ T,G هو (0)
السطر الرابع: استبدال G بـ G هو (1) واستبدال G بـ C,T هو (0)

$$B = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$MSE = 0.9536$$

$$MR = 89.4040$$

2) استبدال القاعدة النتروجينية T والتي عددها 447 من سلسلة الجين MT-ND5 بالرمز C، وتم معالجتها باستخدام الخوارزمية المعدة لهذا الغرض، وكانت النتائج كما يأتي:

مصفوفة الاحتمالات الانتقالية (A) هي:

$$A = \begin{bmatrix} 0.3091 & 0.2400 & 0.3018 & 0.1491 \\ 0.3177 & 0.2282 & 0.3647 & 0.0895 \\ 0.3055 & 0.2894 & 0.3408 & 0.0643 \\ 0.2500 & 0.1719 & 0.4219 & 0.1563 \end{bmatrix}$$

مصفوفة الإصدارات (B) هي:
السطر الأول : استبدال A بـ A هو (1) واستبدال A بـ C,G هو (0)
السطر الثاني : استبدال T بـ C هو (1) واستبدال T بـ A,G هو (0)
السطر الثالث: استبدال C بـ C هو (1) واستبدال C بـ A,G هو (0)
السطر الرابع: استبدال G بـ G هو (1) واستبدال G بـ C,A هو (0)

$$B = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$MSE = 0.2456$$

$$MR = 75.4415$$

واستبدالها بالرمز G، وتم معالجتها باستخدام الخوارزمية المعدة لهذا الغرض، وكانت النتائج كما يأتي:

مصفوفة الاحتمالات الانتقالية (A) هي:

$$A = \begin{bmatrix} 0.3091 & 0.2400 & 0.3018 & 0.1491 \\ 0.3177 & 0.2282 & 0.3647 & 0.0895 \\ 0.3055 & 0.2894 & 0.3408 & 0.0643 \\ 0.2500 & 0.1719 & 0.4219 & 0.1563 \end{bmatrix}$$

مصفوفة الإصدارات (B) هي:

السطر الأول : استبدال A بـ A هو (1) واستبدال A بـ C,G هو (0)
السطر الثاني : استبدال T بـ G هو (1) واستبدال T بـ A,C هو (0)
السطر الثالث: استبدال C بـ C هو (1) واستبدال C بـ A,G هو (0)
السطر الرابع: استبدال G بـ G هو (1) واستبدال G بـ C,A هو (0)

$$B = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$MSE = 0.4238$$

$$MR = 89.4040$$

3) استبدال القاعدة النتروجينية C والتي عددها 622 من سلسلة الجين MT-ND5 بالرمز G، وتم معالجتها باستخدام الخوارزمية المعدة لهذا الغرض، وكانت النتائج كما يأتي:

مصفوفة الاحتمالات الانتقالية (A) هي:

$$A = \begin{bmatrix} 0.3091 & 0.2400 & 0.3018 & 0.1491 \\ 0.3177 & 0.2282 & 0.3647 & 0.0895 \\ 0.3055 & 0.2894 & 0.3408 & 0.0643 \\ 0.2500 & 0.1719 & 0.4219 & 0.1563 \end{bmatrix}$$

مصفوفة الإصدارات (B) هي:

السطر الأول : استبدال A بـ A هو (1) واستبدال A بـ T,G هو (0)
السطر الثاني : استبدال T بـ T هو (1) واستبدال T بـ A,G هو (0)
السطر الثالث: استبدال C بـ G هو (1) واستبدال C بـ A,T هو (0)
السطر الرابع: استبدال G بـ G هو (1) واستبدال G بـ T,A هو (0)

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$MSE = 0.1060$$

$$MR = 89.4040$$

ثانياً: نتائج تطبيق الخوارزمية المقترحة على سلسلة الجين MT-ND5 الخاصة بالفئران

تم اختيار سلسلة الجين MT-ND5 الخاص بالفئران من الموقع MT-Nd5 NADH dehydrogenase 5, mitochondrial [Mus musculus (house mouse)] والتي تتكون من 1822 قاعدة نيتروجينية وذلك لتحديد نوعية القاعدة النتروجينة المستبدلة لسلسة الجين والتي تم الحصول عليها من موقع NCBI ضمن قاعدة بيانات Data Base في مراكز عالمية متخصصة في الهندسة الوراثية ودراسة عمل الجينات، وباستخدام الخوارزمية المقترحة لتحديد نوعية القاعدة النيتروجينية المستبدلة لسلسلة الجين MT-ND5 ، والتي تم برمجتها باستخدام اللغة البرمجية MATLAB R2017b ، وتم استبدال القواعد النتروجينية (A, T, C, G) وكما يأتي:

1) استبدال القاعدة النتروجينية A والتي عددها 625 من سلسلة الجين MT-ND5 بالرمز T، وتم معالجتها باستخدام الخوارزمية المعدة لهذا الغرض، وكانت النتائج كما يأتي:

مصفوفة الاحتمالات الانتقالية (A) هي:

$$A = \begin{bmatrix} 0.3317 & 0.3061 & 0.2532 & 0.1090 \\ 0.3256 & 0.3025 & 0.2813 & 0.0906 \\ 0.3838 & 0.2889 & 0.2646 & 0.0626 \\ 0.3187 & 0.1538 & 0.3297 & 0.1978 \end{bmatrix}$$

مصفوفة الإصدارات (B) هي:

السطر الأول : استبدال A بـ T هو (1) واستبدال A بـ C,G هو (0)
السطر الثاني : استبدال T بـ T هو (1) واستبدال T بـ C,G هو (0)
السطر الثالث: استبدال C بـ C هو (1) واستبدال C بـ T,G هو (0)
السطر الرابع: استبدال G بـ G هو (1) واستبدال G بـ C,T هو (0)

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$MSE = 0.2823$$

$$MR = 71.7738$$

واستبدالها بالرمز C، وتم معالجتها باستخدام الخوارزمية المعدة لهذا الغرض، وكانت النتائج كما يأتي:

مصفوفة الاحتمالات الانتقالية (A) هي:

$$A = \begin{bmatrix} 0.3317 & 0.3061 & 0.2532 & 0.1090 \\ 0.3256 & 0.3025 & 0.2813 & 0.0906 \\ 0.3838 & 0.2889 & 0.2646 & 0.0626 \\ 0.3187 & 0.1538 & 0.3297 & 0.1978 \end{bmatrix}$$

مصفوفة الإصدارات (B) هي:

لسطر الأول : استبدال A بـ T هو (0) واستبدال A بـ C هو (1) واستبدال A بـ G هو (0)
السطر الثاني : استبدال T بـ T هو (1) واستبدال T بـ C,G هو (0)
السطر الثالث: استبدال C بـ C هو (1) واستبدال C بـ T,G هو (0)
السطر الرابع: استبدال G بـ G هو (1) واستبدال G بـ C,T هو (0)

$$B = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$MSE = 1.0675$$

$$MR = 73.3114$$

واستبدالها بالرمز G، وتم معالجتها باستخدام الخوارزمية المعدة لهذا الغرض، وكانت النتائج كما يأتي:

مصفوفة الاحتمالات الانتقالية (A) هي:

$$A = \begin{bmatrix} 0.3317 & 0.3061 & 0.2532 & 0.1090 \\ 0.3256 & 0.3025 & 0.2813 & 0.0906 \\ 0.3838 & 0.2889 & 0.2646 & 0.0626 \\ 0.3187 & 0.1538 & 0.3297 & 0.1978 \end{bmatrix}$$

مصفوفة الإصدارات (B) هي:

السطر الأول : استبدال A بـ T هو (0) واستبدال A بـ C هو (0) واستبدال A بـ G هو (1)
السطر الثاني : استبدال T بـ T هو (1) واستبدال T بـ C,G هو (0)
السطر الثالث: استبدال C بـ C هو (1) واستبدال C بـ T,G هو (0)
السطر الرابع: استبدال G بـ G هو (1) واستبدال G بـ C,T هو (0)

$$B = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$MSE = 0.8995$$

$$MR = 90.0055$$

2) استبدال القاعدة النتروجينية T والتي عددها 520 من سلسلة الجين MT-ND5 بالرمز C، وتم معالجتها باستخدام الخوارزمية المعدة لهذا الغرض، وكانت النتائج كما يأتي:

مصفوفة الاحتمالات الانتقالية (A) هي:

$$A = \begin{bmatrix} 0.3317 & 0.3061 & 0.2532 & 0.1090 \\ 0.3256 & 0.3025 & 0.2813 & 0.0906 \\ 0.3838 & 0.2889 & 0.2646 & 0.0626 \\ 0.3187 & 0.1538 & 0.3297 & 0.1978 \end{bmatrix}$$

مصفوفة الإصدارات (B) هي:

السطر الأول : استبدال A بـ A هو (1) واستبدال A بـ C,G هو (0)
السطر الثاني : استبدال T بـ C هو (1) واستبدال T بـ A,G هو (0)
السطر الثالث: استبدال C بـ C هو (1) واستبدال C بـ A,G هو (0)
السطر الرابع: استبدال G بـ G هو (1) واستبدال G بـ C,A هو (0)

$$B = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$MSE = 0.2510$$

$$MR = 74.9039$$

واستبدالها بالرمز G، وتم معالجتها باستخدام الخوارزمية المعدة لهذا الغرض، وكانت النتائج كما يأتي:

مصفوفة الاحتمالات الانتقالية (A) هي:

$$A = \begin{bmatrix} 0.3317 & 0.3061 & 0.2532 & 0.1090 \\ 0.3256 & 0.3025 & 0.2813 & 0.0906 \\ 0.3838 & 0.2889 & 0.2646 & 0.0626 \\ 0.3187 & 0.1538 & 0.3297 & 0.1978 \end{bmatrix}$$

مصفوفة الإصدارات (B) هي:

السطر الأول : استبدال A بـ A هو (1) واستبدال A بـ C,G هو (0)
السطر الثاني : استبدال T بـ G هو (1) واستبدال T بـ A,C هو (0)
السطر الثالث: استبدال C بـ C هو (1) واستبدال C بـ A,G هو (0)
السطر الرابع: استبدال G بـ G هو (1) واستبدال G بـ C,A هو (0)

$$B = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$MSE = 0.3998$$

$$MR = 90.0055$$

3) استبدال القاعدة النتروجينية C والتي عددها 495 من سلسلة الجين MT-ND5 بالرمز G، وتم معالجتها باستخدام الخوارزمية المعدة لهذا الغرض، وكانت النتائج كما يأتي:

مصفوفة الاحتمالات الانتقالية (A) هي:

$$A = \begin{bmatrix} 0.3317 & 0.3061 & 0.2532 & 0.1090 \\ 0.3256 & 0.3025 & 0.2813 & 0.0906 \\ 0.3838 & 0.2889 & 0.2646 & 0.0626 \\ 0.3187 & 0.1538 & 0.3297 & 0.1978 \end{bmatrix}$$

مصفوفة الإصدارات (B) هي:

السطر الأول : استبدال A بـ A هو (1) واستبدال A بـ T,G هو (0)

السطر الثاني : استبدال T بـ T هو (1) واستبدال T بـ A,G هو (0)

السطر الثالث: استبدال C بـ G هو (1) واستبدال C بـ A,T هو (0)

السطر الرابع: استبدال G بـ G هو (1) واستبدال G بـ T,A هو (0)

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$MSE = 0.0999$$

$$MR = 90.0055$$

والجدول (1) التالي يوضح نتائج عمليات الاستبدال للقواعد النتروجينية الأربع (A, T, C, G) لسلسلة الجين MT-ND5 الخاصة بالإنسان والفئران من خلال متوسط مربعات الخطأ والنسبة المئوية للتطابق لكل عملية استبدال.

الجدول (1): عمليات الاستبدال للقواعد النتروجينية الاربع (A, T, C, G) من سلسلة الجين MT-ND5 الخاصة بالإنسان والفئران.

Original Nitrogenous Base	Substituted Nitrogenous Base	MSE Human	MR Human	MSE Mice	MR Mice
A	T	0.2439*	75.6071*	0.2823	71.7738
	C	1.1236	71.9095	1.0675*	73.3114*
	G	0.9536	89.4040	0.8995*	90.0055*
T	C	0.2456*	75.4415*	0.2510	74.9039
	G	0.4238	89.4040	0.3998*	90.0055*
	A	0.2439*	75.6071*	0.2823	71.7738
C	A	1.1236	71.9095	1.0675*	73.3114*
	T	0.2456*	75.4415*	0.2510	74.9039
	G	0.1060	89.4040	0.0999*	90.0055*
G	A	0.9536	89.4040	0.8995*	90.0055*
	C	0.1060	89.4040	0.0999*	90.0055*
	T	0.4238	89.4040	0.3998*	90.0055*

من الجدول (1) اثبتت الخوارزمية المقترحة في استخدام خوارزمية Viterbi في نموذج ماركوف المخفي انها دقيقة في تحديد نوعية القاعدة النتروجينية المستبدلة لسلسلة الجين ND5-MT الخاصة بالإنسان والفئران وذلك بالاعتماد على النسب العالية للتطابق التي تم الحصول عليها وعلى مجموع مربعات الخطأ المنخفضة. وظهرت من الجدول ان الخوارزمية كانت أفضل في تحديد نوعية القاعدة النتروجينية من سلسلة الجين ND5-MT الخاصة بالفئران مقارنة مع سلسلة الجين الخاصة بالإنسان.

6. الاستنتاجات Conclusions

من النتائج أعلاه تم ملاحظة انه عند استبدال قاعدة نتروجينية معينة (مثل القاعدة A) لسلسلة الجين MT-ND5 الخاصة بالإنسان والفئران مع قاعدة نتروجينية أخرى (مثل القاعدة T)، تم التوصل الى وجود تتطابق بنتائج مجموع متوسط مربعات الخطأ ونسبة التطابق مع عملية الاستبدال بشكل معاكس (أي استبدال القاعدة T بالقاعدة A)، مما يدل على ان دقة تحديد نوعية القاعدة النتروجينة المستبدلة للجين تعتمد فقط على عدد القواعد النتروجينية في سلسلة الجين وليس على النوعية. وتم ملاحظة انه عند استبدال القاعدة النتروجينية G (Guanine) ببقية القواعد النتروجينية تعطي نفس نسبة التطابقMR للإنسان وكذلك للفئران، مما يدل على ان تغيير القاعدة النتروجينية G بقواعد نتروجينية أخرى لا يؤثر على دقة تحديد نوعية القاعدة النتروجينية لسلسلة الجين MT-ND5. كما أظهرت النتائج ان تغيير القاعدة النتروجينية G ببقية القواعد يعطي أعلى نسب للتتطابق، مما يدل على عدم تأثر سلسلة الجين MT-ND5 للإنسان والفئران بالقاعدة النتروجينية G.

References

1. Abdulla,W. H. and Kasabov, N. K. (1999), " The Concept of Hidden Markov Model in Speech Recognition". Dept. of Knowledge Engineering Lab Dept. Information Science, College of Engineering, University of Otage, New Zealand.
2. ChengXiang Zhai (2003). "A Brief Note on the Hidden Markov Models (HMMs)" http://citeseerx.ist.
3. Couvreur.CH. "Hidden Markov Models and Their Mixtures" (1996), Universit_e catholique de Louvain, Facult_e des sciences {D_epartement de math_ematiques.
4. Grant, G.and Ewens, W. (2005), "Statistical Methods in Bioinformatics ". Second Edition, University of Pennsylvania, Philadelphia, USA.
5. Johannesson, P. (1999). "Rain Flow Analysis of Switching Markov Loads", PhD thesis, Lund Institute of Technology, Lund.
6. Mark Stamp (2012)" A Revealing Introduction to Hidden Markov Models". September 28.
7. Oliver C. Ibe "Markov Processes for Stochastic Modeling" (2009), Elsevier Inc. All rights reserved.
8. Robert, J. E; Lakhdar, A; and John B. M. (2008). "Hidden Markov Models Estimation and Control", 3rd printing vol. 29 ISBN 0-387-94364-1, Dept of Systems Engineering, Australia.
9. Teresa, M. P. (2007), " Encyclopedia of the Human Genome: Hidden Markov Models". School of Medicine Johns, Hopkins, USA.
10. Xuan, T. (2004), "Autoregressive Hidden Markov Model with Application in Study ". Thesis of Science, S7N5E6, Dept. of Mathematics and Statistics, University of Saskatchewan, Saskatoon.

Employment of Hidden Markov Model in Determining the Quality of Nitrogenous Base Substituted of MT-ND5 gene Sequence in Humans and Mice

Sura Mohammed Jamal Alden Muthanna Subhi Sulaiman
Department of Statistic and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq

Abstract

Hidden Markov models were developed to analyze bioinformatics data that have attracted the attention of researchers because of their critical importance in the life of living organisms. The aim of this paper was to determine the quality of the nitrogenous base substituted for the MT-ND5 gene chain of humans and mice. The proposed algorithm using the Viterbi algorithm in the Hidden Markov model proved to be good in determining the quality of the nitrogenous base substituted for the MT-ND5 gene chain of humans and mice, depending on the high match ratios obtained and the low sum of squared errors. A computer program was designed for this purpose and the algorithm was programmed in MATLAB R2017b language, and from the practical application of the algorithm it is seen that the Hidden Markov model is a particularly powerful approach to determine the match ratio up to a high classification accuracy.

Keywords: hidden Markov model, Viterbi algorithm, MT-ND5 gene sequence in humans and mice.