مجلة العلوم الاقتصادية والادارية المجلد 13/ع46/لسنة 2007

# استخدام سلاسل ماركوف في التعرف على تعاقبات الحامض النووي DNA

أ.د. ضوية سلمان حسن
قسم الاحصاء – جامعة بغداد

أ.د. علي عبد الرحمن الزعاك
عميد معهد الهندسة الوراثية – جامعة بغداد

زينب هاتف عباس – ماجستير بحوث عمليات

## الملخص

ان للحامض النووي DNA اهمية بالغة لكونهِ حاملاً للمعلومات الوراثيـة. وبالتـالي فـان دراسة تعاقبات الحامض النووي DNA سوف يعطي انطباعاً عن الصفات الوراثية للكائن الحـي. ولاهمية الموضوع ارتأينا تقديم بحث يتضمن التعرف علـى سلاسـل الحـامض النـووي DNA وحساب الاحتمالات الانتقالية الشرطية لعناصر هذه السلاسل، وبناء مصفوفات احتماليـة انتقاليـة من الرتبة الاولى والثانية والثالثةو الرابعة والخامسة وثم التجريب على اربعة تعاقبـات لسلـسلة الدنا لهرمون النمو البشري. وتم الاعتماد على نتائج الطرق البيولوجية ومميـزات الجـين فـي تعاقبات DNA للمفاضلة بين نتائج الطرق الثلاث التي ترشح السلسلة ذات الاحتمالية الاكبر لكي تمثل الجين.

وتم عمل برنامج تشفير وترجمة للنسق لثلاث نظم قراءة، بهـدف المقارنـة بـين تعاقبـات هرمون النمو البشري وسلاسل اخرى.

## Abstract

The DNA has great importance becaûse it is a hereditary information carrier. In view of the importance of markoves chains in explaining and analyzing many of the phenomena through succession, we apply it in our research to obtain the conditional probabilities of chains elements, for its great importance in finding the hereditary genetics from DNA compositions.

Markov chains has been implemented for the probable alignment of the first, second, third, fourth, and fifth orders.

A program was created to code and translate the three models of the ORF for DNA succession and the chains for the three program was written by visual basic language to carry out these operations and to yield sequence alignment between human and growth hormone gene and other chains.

The program implemented DNA chains in which the nucleotide number is less than 3000 bases exemplified by the human growths hormone, mouse, and chicken growth hormone.

## المقدمة

بعد التطور المذهل الذي شهده علم بحوث العمليات والتاثير المتزايد لـه فـي الكثيـر مـن المجالات، اصبحت تطبيقاته واسعة ومتنوعة ولا سيما نتيجة استخدام الحاسوب وقد استخدم حقل OR في مجال الهندسة الوراثية، حيث استخدم الباحثين مواضيع مختلفة من OR مثلاً البرمجـة الديناميكية وسلاسل ماركوف في بحث التعاقبات الجينية المختلفـة لسلـسلة الجينـات المكونـة للحامض النووي DNA، لما لهذه التكرارات من اهمية في ارجاع نقاط التشابه والاختلافات الـى مصادرها الاصلية وكذلك تكوين الشفرات الوراثية ودورها في عملية تكوين الاحمـاض الامينيـة وعملية استنساخها الى البروتين، ودور الحامض النووي في ذلك. تعتبـر الاحنمـالات الانتقاليـة الشرطية بعض الادبيات الخاصة بسلاسل ماركوف، وتعتمد هذه الاحتمالات جنباً الى جنب مـــع المعلومات الحياتـــية وكذلـــك مع نظم القـراءة المفتوحـة (Open Reading Frames) بهدف تشفير وتحليل سلاسل الحامض النووي وايجاد مواقع التـشابه بـين سلـسلتين وتسلـسل النيوكليوتيدات في المناطق المتشابه بين سلسلتين، ولذلك ارتاينا البحث في هذا الموضوع تحقيقـا للهدف التالي.

## هدف البحث

يهدف البحث الى اختيار افضل الطرق المعتمدة على الاحتمالات الانتقالية لمصفوفة ماركوف في ترشيح السلسلة الجينية من السلاسل التي تبدأ بشفرة البداية (Start code) وتنتهي بـشفرة النهاية (Stop code) والمتكونة من انساق النظم المفتوحة (ORF) لسلسلة الحامض النـووي، وذلك من خلال تقدير الاحتمالات الانتقالية لسلاسل ماركوف وتقدير الاحتمالات الانتقالية بطريقـة ماركوف التخمينية لاربعة سلاسل من سلاسل DNA لهرمون النمـو البـشري واختيـار افـضل التقديرات. وتم اعداد برنامج بلغة فيجوال بيسك وهذا البرنامج هو خطوة اولية في عالم البـرامج الجاهزة يساعد الباحثين البايولوجين ويقوم بعملية تشفير وترجمـة لـصيغ القـراءة المفتوحـة للحمض النووي DNA وكذلك عملية المقارنة بين سلسلة النمو البـشري مـع سلـسلة اخـرى لهرمون النمو لاحياء اخرى مثل الدجاج والفئران.

## الجانب النظري

تتكون سلسلة الحامض النووي DNA من تكرار متسلسل لاربعة من القواعد الناتروجينيـة (A,T,G,C) وهي الادنين (A) والثايميين (T) والسايتوسين (C) والكوانين (G). ويكون ارتباط القواعد الناتروجينية بواسطة اواصر هيدروجينية (Hydrogen Bonds) اذ يتصل الادنـين (A) لاحد السلسلتين مع الثايمين (T) للسلسلة الاخرى باصرتين، في حين يتصل السايتوسين (C) في احدى السلسلتين مع الكوانيين (G) للسلسلة الاخرى بثلاث اواصر هيدروجينيـة، اي ان الادنـين (A) هو يقابل الثايمين (T) وان الكوانيين (G) يقابل الـسايتوسين (C) اينمـا تـرد هـذه فـي متسلسلة نيوكليوتيدات الحامض النووي DNA. وتعد هذه اهم الدلائل التي حولت الانظـار الـى الحامض الدنا باعتباره المادة الوراثية، بعد ان كان يعتقد ان كمية البروتين فـي الخليـة تـزداد بصورة مستمرة بينما كمية الدنا فهي ثابتة في الخلية تقريبا، ولكن من اهم صفات الدنا هي قـدرة الدنا على التضاعف اذ تعتمد عملية التضاعف اعتماداً رئيـسياً علـى طريقـة اتـصال القواعـد النتروجينية مع بعضها في الحلزون المزدوج*، كما ذكرنا ان سلسلة الحـامض النـووي DNA تتكون من تكرار متسلسل لاربعة من القواعد الناتروجينية (A,T,G,C) والاختلاف الحاصل فـي

لاخر، اي ان الموديل الذي يمثل ترتيب هذه القواعد الاربعة والانتقال من نيوكلتيد الى الاخر يتمثل بالزمن، وافضل ما يعبر عن هذا الموديل بالعمليات التصادفية، والتي تمثل على شكل سلاسل ماركوف، لكي يسهل علينا حساب تكرارات القواعد لكل جين، وحساب الاحتمالات الانتقالية للانتقال من نيوكلتيد الى اخر، تكون بطريقة الامكان الاعظم.

## تستخدم نماذج ماركوف التخمينية Interpolated Markov Models (IMM)

للزيادة في دقة المعلومات التي نريد الحصول عليها في نماذج ماركوف ذات الرتب العالية اذ ان هناك عدد كبير من الثوابت للرتب العالية (ان عدد ثوابت الرتبة يتناسب طردياً مع رتبة الموديل) وتعتمد هذه السلاسل على الوزن التخميني ويتم بواسطة IMM حساب القيمة المتوقعة التخمينية لكل سلسلة جزئية من السلسلة لكل عنصر من عناصر السلسلة كعلاقة رياضية خطية من الاحتمالات الانتقالية لرتب مختلفة من سلاسل ماركوف اذ ان :

$$E_X(x_i \mid x_{i-1}, x_{i-2}, \ldots\ldots x_{IMMk}) = \sum \lambda_{j,i} \Pr(x_i \mid x_{i-1}, x_{i-2}, \ldots\ldots x_{IMMk}) \quad (1)$$

حيث ان :

$\Pr(x_i \mid x_{i-1}, x_{i-2}, \ldots\ldots x_{IMMk})$ يمثل القيمة الاحتمالية الانتقالية لمصفوفة ماركوف $\lambda_{j,i}$ ثابت يمثل الوزن لقيم الاحتمالات الانتقالية لموديل من الرتبة $j_{th}$ .

ان الهدف من هذا الموديل، هو استعمال الاوزان في الموديلات ذات الرتب العالية اذ ان الموديل من الرتبة $K_{th}$ يكون عدد الثوابت كبيراً فيه وتكون السلسلة طويلة لذلك فان تكراره يكون قليلاً وبالتالي وزن هذا الحد يكون اكبر مما يعطي فرصة اكبر للرتب العالية. تستعمل الاحتمالات من الرتب الدنيا التي تمتلك حدوث عالٍ (تكرار عالي) للمساعدة في تصحيح الاخطاء التي تقع عند تقدير القيمة الاحتمالية الانتقالية للرتب العليا من موديلات ماركوف. لذلك فان استعمال اعداد مختلفة من الثوابت يمكن السيطرة على حالات مختلفة لحساب الاحتمالات الانتقالية. وهناك معايير مختلفة لحساب قيم الثوابت او الاوزان التقديرية لموديل ماركوف منها طريقة ماركوف التخمينية $IMM_k$ وطريقة $X^2$ والطريقة الافتراضية وغيرها (وقد فصلت هذه الطرائق في المصدر 4)* من المصادر العربية المشار اليها في نهاية البحث.

اضافة للطرق السابقة يوجد معيار اخر لحساب قيمة الوزن $\lambda i$ للسلسلة الجزئية في موديلات ماركوف التخمينية فتعتمد على طريقة افتراضية تتمثل باعطاء اوزان متساوية ($\sum \lambda_i = 1$) لجميع الرتب.

$$E_X(X_i \mid X_{i-1}, X_{i-2}, \ldots\ldots X_{i-n}) = \lambda_0 P(X_i) + \lambda_1 P(X_i \mid Xi\text{-}1) + \lambda_n P(X_i \mid X_{i-1}, \ldots\ldots X_{i-n}) \quad (2)$$

فعلى سبيل المثال لو كان لدينا موديل من الرتبة الرابعة ونريد ايجاد القيمة الاحتمالية التخمينية له، فان الوزن يقسم بالتساوي على عدد الحدود، لو كان عندنا السلسلة الجزئية (atca) فان احتمال ان تاتي القاعدة g بعد السلسلة الجزئية اي ان $\Pr(g \mid atca)$

$$IMM\ 4 = \lambda 0\ p(a) + \lambda 1\ p(c|a) + \lambda 2\ p(t|ca) + \lambda 3\ p(a|tca) + \lambda 4\ p(g|atca)$$

$$\lambda 0 = \lambda 1 = \lambda 2 = \lambda 3 = \lambda 4 = 1/5 \quad (3)$$

ولكي نمثل سلسلة DNA بطريقة ماركوف نستخرج الاحتمال الانتقالي لكل قاعدة نتروجينية وذلك بايجاد التوقع الاحتمالي من الدرجة الاولى فمثلاً احتمال ان ياتي الادنين في الموقع الثالث من الشفرة فقط يعتمد على وجود الكوانيين في الموقع الثاني من الشفرة، ويعني ذلك رياضياً

Pr ( a3|g2)

ولايجاد القيمة الاحتمالية للسلاسل الجزئية المكونة لسلسلة الدنا (S) نطبق التعريف التالي

$$Pr(S) = Pr(X_L, X_{L-1}, X_{L-2}, \ldots\ldots X_1)$$

$$= Pr(X_L | X_{L-1}, X_{L-2}, \ldots X_1)\, Pr(X_{L-1}|X_{L-2}, \ldots X_1)\, Pr(X_1) \qquad (4)$$

$$= Pr(X_L | X_{L-1}, X_{L-2})\, Pr(X_{L-1}|X_{L-2}, X_{L-3}) \ldots Pr(X_3|X_2, X_1)\, Pr(X2|X1)\, Pr(X_1)$$

ان مخرجات هذه السلسة تعتمد على مجموعة الاحتمالات الانتقالية لسلسلة ماركوف .

## الجانب التطبيقي

تم اعتماد هرمون النمو البشري كاساس في التطبيق العملي وقد اخذت امثلة من بدائية النواة بالاعتماد على تعاقبات (الذيفانات المعوية) Enterotxins المنتجة من قبل بكتريا القولون الشيكلا والسالمونيلا وكذلك تم اعتماد متسلسلات هرمون النمو للدجاج، البقر والفئران لغرض التطبيق والمقارنة مع هرمون النمو البشري الذي تفرزه الغدة النخامية متوجهاً الى بقية الاعضاء معطياً الاشارة لافراز عوامل النمو الاخرى لتبدأ دورها في حث خلايا الجسم على الانقسام والتكاثر والتمايز.

ومن المعلوم ان الذيفانات المعوية هي انزيمات (سموم) تفرزها انواع عديدة من بكتريا مثل (E.coli , Shiglla, Salmonella) داخل الامعاء وهذه السموم المفرزة من البكتريا تمتصها الخلايا المبطنة للامعاء ، ولذا يكون تاثير هذه السموم موضعي وبعيد عن مصدرالافراز وهذا هو سبب انتقال السموم بواسطة الدم . وتختلف الذيفانات في تركيبها باختلاف الجينات المشفرة لها ولذلك اصبح من الضروري دراسة هذه الجينات لمعرفة الاساس الجزيئي للامراض البكترية .

تضمن التطبيق العملي للبحث اولا كتابة برنامج بلغة في جوال بيسك يقوم بتشفير Coding لمتسلسلة الحامض النووي DNA وكذلك عملية ترجمة Translation للبروتين لمتسلسلة DNA مع انساق القراءة المفتوحة ORF ، وكذلك مقارنة السلاسل (Sequence alignment) لسلسلتين من الحامض النووي الاولى لهرمون النمو البشري والاخرى لهرمون النمو من الحيوانات كالابقار والدجاج والفئران .وايجاد مناطق التشابه بين السلسلتين ولتمييز اكثر الكائنات تشابهاً مع الهرمون النمو البشري وتضمن التطبيق ثانياً حساب الاحتمالات الانتقالية الشرطية من سلسلة الحامض النووي DNA حيث تم تطبيق سلاسل ماركوف وبناء مصفوفة الاحتمالات الانتقالية من الرتبة الاولى والثانية والثالثة والرابعة والخامسة وفق خوارزمية اعدت لهذا الغرض ، فضلاً عن برنامج تشفير وترجمة لكل من متسلسلات الحامض النووي. وقد اضيفت للبرنامج واجهات مرئية تسهل بناء واجهة للبرنامج تتيح للباحث سهولة الاستعمال والذي يقوم بعمليتين الاولى تتمثل بعملية تشفير (Coding) والترجمة (Translation) لانساق (ORF) لـ DNA ، وذلك لتحديد مناطق التشابه وتسلسل الشفرات بين السلسلتين ويتم تحديدها من قبل المستخدم من السلاسل التي حمل بها البرنامج وقد تضمن برنامج Genetics الذي تم اعداده في هذا البحث سلاسل جينية لهرمون النمو لعدة كائنات حقيقية النواة مع السلاسل (الذيفانات) بدائية النواة والتي عدد القواعد الناتروجينية بها اقل من 3000 قاعدة ناتروجينية ، وقد حمل البرنامج بـ 22

وفي التطبيق العملي، حسبت الاحتمالات الانتقالية لمصفوفة ماركوف للتراكيب الجينية لسلاسل DNA لكل قاعدة نتروجينية او لكل مقطع جزئي من السلسلة بطريقة الامكان الاعظم وتتمثل الرتبة Zero القيم الابتدائية للقواعد النانروجينية في السلسلة، اي مدى احتمالية وجود القاعدة النانتروجينية في السلسلة لهرمون النمو البشري ووجد ان:

P0 (a) = 0.20281
P0 ( c ) = 0.26506
P0 (g) = 0.30723
P0 (t) = 0.22490
P(S) = P(t) * p(t) P( c ) P(t) .....P(a)P(g)P(g)

ووجد ان القيمة الاحتمالية الابتدائية لطريقة ماركوف الثابتة $FMM_0 = 7.98385\ E^{-298}$

اما الرتبة الاولى والتي تتمثل باعتماد القاعدة النتروجينية المراد ايجاد الاحتمالات الانتقالية لها على القاعدة التي تسبقها والمصفوفة التالية تمثل الاحتمالات الانتقالية لهرمون النمو من الرتبة الاولى:

| | A | T | C | G |
|---|---|---|---|---|
| A | 0.28712 | 0.11882 | 0.12872 | 0.46534 |
| T | 0.10715 | 0.26786 | 0.28571 | 0.33928 |
| C | 0.20455 | 0.34848 | 0.37128 | 0.07575 |
| G | 0.21568 | 0.15688 | 0.24836 | 0.37908 |

وبتطبيق القانون

Pr(S) = P(t)P(t | t) P(t | c) P(c | c) P(c |t) .....P(t|g)

نحصل على القيمة الاحتمالية الكلية للسلسلة من الرتبة الاولى لطريقة ماركوف ذو الرتبة الثانية التي تساوي $9.9727E^{-280}$

اما الرتبة الثانية والتي تتمثل بمدى الاحتمال تاتي بها القاعدة النتروجينية بعد القاعدتين التي تسبقها والتي تتمثل مثلاً (TT) وهكذا تم حساب الاحتمال الكلي لسلسلة هرمون النمو البشري باعتماد الاحتمالات الانتقالية من الرتبة الاولى والثانية والثالثة والرابعة والخامسة لطريقة (FMM) واوضحت جميع هذه النتائج في المصدر رقم 4 من المصادر العربية. وكذلك اعطيت نتائج المقارنة بين طريقة ماركوف التخمينية المستندة على الاوزان الثابتة وتلك المستندة على الاوزان المستخرجة بطريقة كاي سكوير كما هو مفصل في الرسالة.

## الاستنتاجات

1. عند تطبيق برنامج Genetics وجد ان التركيب الجيني لهرمون النمو البشري يشابه الى حد كبير التركيب الجيلي لهرمون النمو البشري للفئران.
2. وبتطبيق طرق ماركوف على السلاسل الجينية لوحظ انه كلما زادت رتبة الموديل (كلما كبرت السلسلة الجزيئية) كلما زادت القيمة الاحتمالية للسلسلة ممايعني ان الرتب العليا هنا الخامسة هي الافضل في التطبيقات البايولوجية .
3. من التطبيق العلمي للبرنامج نلاحظ ان عملية المقارنة (المطابقة) بين سلسلتين وعلى اساس حامضين امينين (اي ستة من القواعد النتروجينية) بان نسبة عالية من التشابة بين هرمون النمو البشري وسلاسل هرمون النمو الفئران وكذلك لسلاسل الذيفانات المنتجة من قبل بكتريا E.coli ، اما عندما تستند على ثلاث حوامض امينية فاكثر فان عملية التشابه تقل كلما ازداد عدد الشفرات المستندة في اساس عملية التشابه .

## التوصيات

1. ان التوسع في البرامج التطبيقية للتراكيب الجينية بلغات متطورة لكي تكون سهلة المنال بالنسبة للباحث البايولوجي ولكون اكثر البرامج التطبيقة تتناول تطبيقات على سلاسل DNA من عمل الباحثين اجانب وهي ذات كلفة عالية.
2. نوصي بتوسع البرنامج لكي يلائم تطبيقات هندسية اعتماداً على الاحماض الامينية صعوداً نحو اكتشاف الدنا المشفر لها.

## المصادر العربية

1. البكري، غالب حمزة (1991)، مبادئ الهندسة الوراثية كتاب منشور عن وزارة التعليم العالي والبحث العلمي – جامعة البصرة – كلية العلوم .
2. الزيدي، عمار الكوني ناصر (2004) التحليل الطيفي لمتسلسلة الحامض النووي DNA – اطروحة دكتوراه مقدمة الى كلية الادارة والاقتصاد – الجامعة المستنصرية .
3. الشيخلي، فؤاد عبد الرزاق (2001) ،" تنفيذ وتقويم الموذج قواعد معلومات الاستنباطية المستندة على جزيئات الحامض النووي DNA اطروحة دكتوراه مقدمة الى قسم علوم الحاسبات الجامعة التكنولوجية.
4. الركابي، زينب هاتف عباس (2005)، استخدام سلاسل ماركوف في التعرف على تعاقبات الحامض النووي DNA – رسالة ماجستير مقدمة الى قسم الاحصاء – جامعة بغداد.

## المصادر الانكليزية

1. Burge , C.L . Karlin , S.(1997) " Prediction of Complete Gene Structures in Human Genomic DNA " J . Mol.Bio. Vol.268 , PP 78 – 94.
2. Delmas , J (2002) " An Introduction to DNA Analysis using Statistical Tolls" Bioinformatics , Vol.33, June . PP – 301-306 .
3. Irving Robent W.(2004) " Plagiarism and Collusion Delection using The Simth – Waterman Algorithm.