استخدام نماذج ماركوف المخفية في المعلوماتية الحيوية

أ.د. عبدالرحيم خلف راهي * اسماء حسين سمير **

المستخلص :

انصب اهتمام هذا البحث على توظيف نماذج ماركوف المخفية Hidden Markov models(HMMs) في المعلوماتية الحيوية Bioinformatics ، لتحديد متسلسلة الحامض النووي DNA للانسان والمتكون من سلسلة من القواعد النيتروجينية التي يطلق عليها النكليوتيدات Nucleotides وهي الادنين Adenine ويرمز له بالحرف a ، السيتوسين Cytosine ويرمز له بالحرف c ، الثايمين Thymine ويرمز له بالحرف t ، الكوانين Guanine ويرمز له بالحرف g ، تم التعامل مع مشكلة تحليل متسلسلة الحامض النووي DNA كمشكلة لغوية تتكون من اربعة حرف هي (a , t , c , g) والتنبؤ بمتسلسلة الحامض النووي DNA باستخدام خوارزمية فيتربي (Veterbi) وتم التوصل الى ان متسلسلة المتنبئ كانت مطابقة لمتسلسلة الحامض النووي DNA للجين CCR5 delta-24 allele .

الكلمات الاساسية : نماذج ماركوف المخفية ، خوارزمية فيتربي ، المعلوماتية الحيوية

Using Hidden Markov Models in Bioinformatics

Abstract :

This research focused on the recruitment of Hidden Markov models (HMMs) in Bioinformatics, to identify DNA sequence of humans and consisting of a sequence of nitrogenous bases, called nucleotides, Adenine are signified by the letter a, Cytosine and signified by the letter c, Thymine and signified by the letter t, Guanine signified by the letter g, is dealing with the problem of analyzing DNA sequence DNA as a linguistic problem consists of four letters (a , c, t, g) and Ltinaya a series of DNA using an algorithm Veterbi .

Keywords: Hidden Markov models (HMMs), Veterbi Algorithm, Bioinformatics .

المبحث الاول

المعلوماتية الحيوية Bioinformatics

1-1 المقدمة : [3] [7]

ان التطور الهائل في علم الاحياء ادى الى انتاج قواعد بيانات ضخمة تتعلق بتركيب الحامض النووي DNA ، تركيب الموروثات ، وظائف البروتينات . تم توظيف التقنيات الالكترونية والبرامجيات بما يتلاءم ونوع هذه البيانات للحفاظ عليها والاستفادة منها ، ان وجود الانترنيت ساعد ان تكون هذه البيانات متاحة للدارسين والباحثين .

* الجامعة المستنصرية / كلية الادارة والاقتصاد .
** باحثة .

تأريخ استلام البحث 2015/12/15
تأريخ قبول النشر 2016/1/19
مستل من اطروحة دكتوراة

هذا التطور ادى الى نشوء علم المعلوماتية الحيوية Bioinformatics وهو علم جديد يجمع بشكل اساس ما بين علم الحاسبات وعلم الاحياء والذي يعرف على انه ادارة المعلوماتية الحيوية بأستخدام التقانة الحاسوبية والمعلوماتية ، كذلك عرف المركز الوطني لمعلومات التكنولوجيا الحيوية (National (NCBI Center for Biotechnology Information علم المعلوماتية الحيوية "العلم الذي نتج عن اندماج علم الحاسب الالي والاحياء وتكنولوجيا المعلومات لتكوين مجال علمي واحد " أستخدمت المعلوماتية الحيوية على نطاق واسع في ابحاث الجينوم البشري لتحديد سلسلة الحامض النووي DNA الكاملة للانسان ، كما ان احد مهام هذا العلم والتي تمثل التحدي الاكبر لحد هذه اللحظة هي مشكلة ايجاد الجين Gene finding في متسلسلة الحامض النووي DNA ، يعتمد علم المعلوماتية الحيوية Bioinformatics على علوم الرياضيات والحاسوب والاحصاء والطب والكيمياء للاستفادة من هذه العلوم في تحليل البيانات .

تهدف المعلومات الحيوية الى الاتي :-

1. استحداث اساليب تقنية وبناء خوارزميات وبرامج تساهم في الحصول على المعلومات من مجموعة ضخمة من البيانات .
2. تحليل وتفسير الانواع المختلفة من البيانات التي تتضمن سلاسل الاحماض الامينية والقطع والبنى البروتينية .
3. تحليل وتفسير بيانات المورثات البشرية .
4. تطوير وتنفيذ ادوات تساعد على ادارة فعالة للانماط المختلفة في المعلومات .
5. الحصول على معلومات جديدة تخص ترميز البروتينات ووظيفة البروتينات والعديد من الوظائف الحيوية الاخرى من مجموعة البيانات الخام ، هذه المعلومات الجديدة مطلوبة لتصميم الادوية والتشخيص الطبي والعلاج الطبي فضلاً عن الميادين البحثية الاخرى .

ان المجالات البحثية للمعلوماتية الحيوية كثيرة منها على سبيل المثال :

1- تقصي الجين Gene finding

وهو معرفة الجينات على سلسلة الحامض النووي DNA وكذلك التنبؤ بتركيب الجين اي تحديد الاكسونات Exons (وهي المناطق المشفرة Coding Sites) ومناطق اللصق (.splice sites) ، والانترونات Entrons (وهي مناطق غير مشفرة non-Coding Sites) .

2- محاذات السلاسل Sequence Alignment

مقارنة السلسلة التي نحصل عليها مع سلسلة او سلاسل اخرى معرفة سابقا فاذا كان هناك تشابه فهذا يعني وجود وظيفة مشتركة او مشابهة بمعنى اخر ان السلسلة الجديدة تؤدي نفس الوظيفة .

3- تركيب البروتينات Protein foldin

معرفة التركيب الثنائي والثلاثي والرباعي (ان وجد) للبروتينات لمعرفة وظيفة البروتين وذلك لوجود ارتباط وثيق بين شكل البروتين ووظيفته .

4- تحديد مواقع ارتباط عوامل النسخ Transcription Factor Binding Site Identification

وهي مواقع صغيرة جدا على سلسلة الحامض النووي DNA ، ان اغلب البيانات الخام المستخدمة في المعلوماتية الحيوية Bioinformatics هو عبارة عن متسلسلة من النكليوتيدات Nucleotides التي تمثل الهيكل الاساسي للحامض النووي DNA ، ومتسلسلة من الاحماض الامينية المناظرة للهيكل الاساسي لليروتينات .

2-1 الهدف :

توظيف نماذج ماركوف المخفية Hidden Markov models(HMMs) في علم المعلوماتية الحيوية Bioinformatics ، لتحديد متسلسلة الحامض النووي DNA .

المبحث الثاني

نماذج ماركوف المخفية Hidden Markov Models (HMMs)

2-1 المقدمة :

تعد نماذج ماركوف المخفية Hidden Markov Models (HMMs) تعميما لنماذج ماركوف الاعتيادية وتطويرا لعمليات ماركوف ، ففي نماذج ماركوف الاعتيادية Markov Models تكون الحالات مرئية بشكل مباشر لذا تكون احتمالات انتقال من حالة الى حالة اخرى معلوم بينما في نماذج ماركوف المخفية (HMMs) تكون الحالات غير مرئية .

ان نماذج ماركوف المخفية (HMMs) هي عبارة عن مجموعة منتهية من الحالات كل حالة تقترن بتوزيع احتمالي ، بشكل عام تتولد الحالة الناتجة طبقا للاحتمالات المقترنة بالحالة حيث توجد احتمالات ناتجة فقط ولاتوجد حالة ظاهرة يمكن مشاهدتها ، لذا تكون الحالات مخفية وهذا هو مفهوم نماذج ماركوف المخفية (HMMs) بشكل عام .

تعد نماذج ماركوف المخفية Hidden Markov Models (HMMs) اداة احصائية قوية للتنبؤ بسلسلة الحالة من خلال سلسلة المشاهدات ، تم الاستفادة منها وتطبيقها في علم المعلوماتية الحيوية Bioinformatics التي تهتم بقواعد البيانات الحيوية والوراثية وادارتها وتطويرها .

2-2 معلمات نماذج ماركوف المخفية (HMMs) [4] [5]

وان معلمات نماذج ماركوف المخفية (HMMs)هي كالاتي :-

1- مصفوفة الاحتمالات الانتقالية ويرمز لها D(nxn)

يرمز للعنصر في الصف i والعمود j بالرمز dij وهو يمثل احتمال الانتقال من الحالة i الى الحالة j ويجب ان يحقق الشرط الاتي :-

$$\sum_{i=0}^{n} d_{ij} = 1 \text{---------- (1)}$$

Where i=1----------n ,j=1----------n

بالنسبة لمتسلسلة الجين التي سيتم دراستها تتكون من اربع حالات هي (a , t , c, g) اي ان n=4

2- مصفوفة العلاقات (الاحتمالات المقترنة بالحالة) ويرمزلها Z(nxm)

يرمز للعنصر في الصف i والعمود k بالرمز Zik وهو احتمال رمز المشاهدة للمؤشر k المنبعث من الحالة i ويجب ان يحقق الشرط الاتي

$$\sum_{k=0}^{m} Z_{ik} = 1 \text{---------- (2)}$$

Where i=1----------n , k=1----------m

في نموذج البحث هناك اربعة احتمالات مقترنة في كل حالة اي ان m=4

3- متجه التوزيع الابتدائي ويرمز له بالرمز ∏ ويجب ان يحقق الشرط الاتي :-

$$\sum_{i=0}^{n} \pi ij = 1 \text{ ---------- (3)}$$

ويعبر عن الانموذج (D, Z , ∏) =λ

2-3 المسائل الاساسية لنماذج ماركوف المخفية [4] [5]

1- مسألة التقويم (Evolution Problem)

وهي عملية ايجاد P(O/λ) اي ايجاد الامكان الاعظم لمتسلسلة المشاهدات $O=(o_1,o_2\text{----------}o_T)$ عندما يكون الانموذج (D, Z , ∏) =λ

2- مسالة حل الشفرة (Decoding problem)

وهي عملية ايجاد متتابعة الحالة المثلى للانموذج (D, Z , ∏) =λ ومتسلسلة المشاهدات $O=(o_1,o_2\text{----------}o_T)$

3- مسألة التدريب (Training Problem)

وهي عملية تعديل معلمات الانموذج (D, Z , ∏) = λ لتعظيم P(O/λ)

ويتم حل هذه المسائل من خلال الخوارزميات الاتية

- الخوارزمية الامامية الخلفية (Forward-Backward Algorithm)
- خوارزمية فيتربي (Viterbi Algorithm)
- خوارزمية بوم ولتش (Beaum-Welch)

عند توظيف نماذج ماركوف المخفية في المعلوماتية الحيوية فان الزمن Time للمتسلسلات الحيوية يعني مواقع النكليوتيدات على طول متسلسلة الحامض النووي DNA .

2-3-1 الخوارزمية الامامية الخلفية (Forward-Backward Algorithm)

اولاً : الخوارزمية الامامية Forward-Algorithm

تعرف الاحتمالية الامامية $\alpha_t(i)$ او ما يدعى (α-Pass) على انها الاحتمالية لمتسلسلة المشاهدات الجزئية والحالة $(q_t=S_i)$ عندما يكون المعطى هو النموذج (λ) والتي يعبر عنها بالصيغة الاتية :-

$$\alpha_t(i)=P(O_1,O_2\text{----------}O_t,q_t=S_i/\lambda)$$

ان هذه الدالة الاحتمالية يمكن ان تحل لـ(N) من الحالات و (T) من المشاهدات يشكل تكراري كالاتي :-

1- البداية (Initialization)

$$\alpha_t(i)=\Pi_i * b_i(O_1) \qquad i=1,2\text{--------}N$$

2- التعاقب (Induction)

$$\alpha_t(j)=[\sum_{i=1}^{N} \alpha_{t=1}(i)a_{ij}]*z_{ij}(O_t)$$

$$J=1,2,\text{----------},N \qquad t=2,3,\text{--------}T$$

3- النهاية (Termination)

$$P(O/\lambda)=\sum_{i=1}^{N} \alpha_{T=1}(i)$$

ثانياً : الخوارزمية الخلفية Backward Algorithm

يعرف المتغير الخلفي $\beta_t(i)$ او مايدعى (β-Pass) على ان الاحتمالية لمتسلسلة المشاهدات الجزئية $O_{T+1},O_{T+2},\text{--------}O_T$ عندما يكون المعطى هو الحالة (S_i) عند الزمن (t) والنموذج (λ) والذي تم التعبير عنه بالصيغة الاتية :-

$$\beta_t(i)=P(O_{T+1},O_{T+2},\text{--------}O_T/q_t=S_i)$$

ان هذا الاجراء مشابه للاجراء الامامي لكن يكون سريان الحالة بالاتجاه الخلفي فيأخذ المشاهدة الاخيرة في اللحظة الزمنية (T) لحين الوصول الى اللحظة الاولى وكالاتي :-

1 ⟵ - - - - - - - - ⟵ T-2 ⟵ T-1 ⟵ T

ان هذه الدالة الاحتمالية يمكن ان تحل لـ (N) من الحالات و (T) من المشاهدات بشكل تكراري كالاتي :-

1- البداية (Initialization)

$$\beta_T(i)=1 \qquad i=1,2,\text{----------}N$$

2- التعاقب (Induction)

$$\beta_t(i)=\sum_{j=1}^{N} d_{ij}z_j(O_{t+1})\beta_{t+1}(j) \qquad j=1,2,\text{----------},N \qquad t=T-1,T-2,\text{----------},1$$

3- النهاية Termination

$$P(O/\lambda)=\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)$$

2-3-2 خوارزمية فيتربي Vfeterb Algorithm

هي طريقة البرمجة الديناميكية Dynamic Programming وتعمل على ايجاد مسار الحالة الانتقالية الاكثر احتمالا عندما يكون المعطى النموذج $\lambda(D,Z,\Pi)$ وعدد الحالات هو N ومتتابعة المشاهدات $O=(o_1,o_2,\text{--------}$ $o_T)$, ذات الطول T وان المتغير الاساسي لهذه الخوارزمية هو المتغير $\delta_t(i)$ حيث يمثل اعلى احتمالية على طول المسار الوحيد في الحالة (i) عندالزمن (t) والذي يساوي متسلسلة الحالة الجزئية الاكثر احتمالا بالنسبة لمتسلسلة المشاهدات المنتهية في الحالة (i) ، وهذه الخوارزمية مشابهه للخوارزمية الامامية الا انها تاخذ اعلى Max احتمالية للمسارات السابقة بينما الخوارزمية الامامية تاخذ مجموع Sum احتمالية المسارات السابقة، كذلك خوارزمية فيتربي لها مؤشرات ترجعية (Back-Pointers) لاتملكها الخوارزمية الامامية ، ويتم ايجاد التسلسل الامثل للحالة وذلك بالاحتفاظ بمسار الحالات المخفية التي تقود لكل حالة ،ثم تتبع افضل مسار (Back-Trace) الى البداية.

اما خطوات خوارزمية فيتربي فهي كالاتي :

1- البداية Initialization

$\delta_1(j)= \pi_i z_j(O1)$, j = 1,2,-------,N $\Psi_1(j)=0$

حيث ان $\Psi_t(j)$ هو متغير حفظ تتبع الاثر (keep track) .

2- التعاقب Induction

$\delta_t(j)=Max^N_{i=1}[\delta_{t-1}(i)d_{ij}]z_j(O_t)$ j=1,2,------,N , t=2,3,------,T

$\Psi t(j)=arg\ Max^N_{i=1}[\delta t-1(i)z_{ij}]$ j=1,2,------,N , t=2,3,------,T

3- النهاية Termination

$P^*=Max^N_{i=1}[\delta_T(i)]$
$q^*=arg\ Max^N_{i=1}[\delta_t(i)]$

4- التعاقب المعاكس Back Tracking

$q^*_t=\Psi t+1(q^*_{t+1})$ t=T-1,T-2,--------,2,1

ثم تقرأ المتسلسلة الافضل للحالات من متجهات (Ψ_t) ولهذا فان خوارزمية فيتربي تؤدي الى نتيجتين مفيدتين هما :-

1- يكون الاختيار (من بين كل المسارات) هو المسار الافضل $Q^*=[q_1^*,q_2^*,------q_T^*]$ والذي يطابق متسلسلةالحالة التي تعطي الامكان الاعظم لمتسلسلة المشاهدات .
2- الامكان على طول المسار الافضل يكون $P(O,Q^*/\lambda)=P^*(O/\lambda)$.

بالمقارنة بالخوارزمية الامامية فان خوارزمية فيتربي تقوم بحساب الامكان على طول المسار الافضل .

3-3-2 خوارزمية بوم ولتش (Beaum-Welch)

تستخدم هذه الخوارزمية لاعادة تقدير معلمات النموذج $\lambda(D,Z,\prod)$ من اجل تعظيم الاحتمالية لمتسلسلة المشاهدات وذلك بالاعتماد على المتغيرين الاتيين :-

المتغير الاول :-

هو $Y_t(i)$ وهو يمثل كون العملية في الحالة (S_i) عند الزمن (t) عندما يكون المعطى هو المشاهدات والنموذج ، ويعبر عنه كالاتي

$$Y_t(i)=\frac{P(q_t=S_i,O/\lambda)}{P(O/\lambda)}$$

$$Y_t(i)=\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}$$

المتغير الثاني :-

هو $\pounds_t(i,j)$ وهو يمثل كون العملية في الحالة (S_i) عند الزمن (t) والحالة (S_j) عند الزمن (t+1) ، ويعبر عنه كالاتي :-

$$\pounds_t(i,j)= \frac{P(q_t=S_i,q_{t+1}=S_j/O,\lambda)}{P(O/\lambda)}$$

$$\pounds_t(i,j)= \frac{\alpha_t(i)d_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\ \sum_{j=1}^{N}\alpha_t(i)d_{ij}b_j(O_{t+1})\beta_{t+1}(j)}$$

2-4 استخدام (HMMs) في المعلوماتية الحيوية [2] [3] [7]

منذ ان تم اعتماد نظرية ماركوف المخفية بصورة رسمية في اواخر الستينيات القرن الماضي ، ان الكثير من العلماء طبق النظرية في المشاكل اللغوية والكلام لكونها مناسبة للمشكلة ، حيث وجود كميات كبيرة من البيانات والمعرفة المستمدة من الملاحظة وهذا هو الحال ايضا في الجانب الحيوي .

استخدمت المعلوماتية الحيوية على نطاق واسع في ابحاث الجينوم البشري لتحديد سلسلة الحامض النووي DNA للانسان والذي هو عبارة عن سلسلة من القواعد النيتروجينية التي يطلق عليها النكليوتيدات Nucleotides وهي الادنين Adenine ويرمز له بالحرف a ، السيتوسين Cytosine ويرمز له بالحرف c ، الثايمين Thymine ويرمز له بالحرف t ، الكوانين Guanine ويرمز له بالحرف g . اثنان من اهم مشاكل تحليل التسلسل الاكثر دراسة هي :-

1- تمييز الكلام Speech Recognition
2- معالجة اللغة Language Processing

ان سلسلة الحامض النووي DNA هي مماثلة لمشكلة تمييز الكلام Speech Recognition ومعالجة اللغة Language Processing وهي تمثل أنموذج الايسر الايمن ، لذلك يمكن التعامل مع مشكلة تحليل متسلسلة الحامض النووي DNA كمشكلة لغوية تتكون من اربعة احرف هي (a , t , c . g) ، فمثلا مشكلة ايجاد الجين Gene Finding في الحامض النووي DNA كتحليل اللغة الى كلمات وجمل ذات معنى وعليه يمكن ان نتعامل بنفس التقنيات المستخدمة للتعرف على الكلام ومعالجة اللغة . مثلما استخدمت نظرية HMM كنماذج رياضية للغة يمكن استخدامها كذلك كنماذج رياضية لتحليل متسلسلة الحامض النووي .

ان النهج القائم على اساس (HMMs) هو نهج مناسب للمشاكل الحيوية الاتية :-

1- محاذات السلاسل Sequence Alignment
وهي مشكلة متكررة في المعلوماتية الحيوية .

2- رسم الخرائط الوراثية Genetic Mapping
وهو من اول تطبيقات (HMMs) في المعلوماتية الحيوية وهو تقدير بعض انواع المسافات الفاصلة بين المواقع الوراثية على طول الكروموسوم .

3- التنبؤ بالتركيب الثانوي للبروتين Secondary Structure Protein Prediction
تعمل (HMMs)على التنبؤ بالهيكل الثانوي للبروتين وهي خطوة مهمة للتنبؤ بهيكل ثلاثي الابعاد.

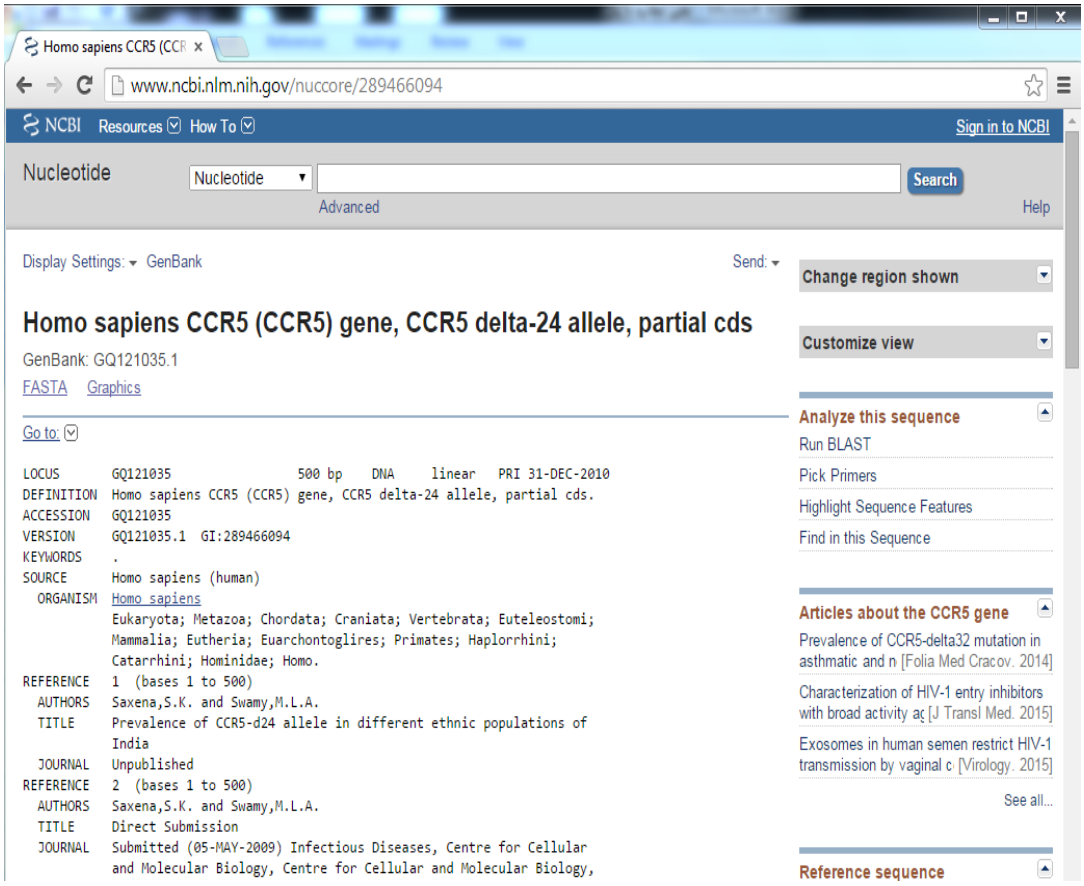
4- ايجاد الجين Gene finding
وهو كما بين سابقا معرفة الجينات على سلسلة الحامض النووي DNA وكذلك التنبؤ بتركيب الجين اي تحديد الاكسونات Exons والانترونات Entrons .

المبحث الثالث

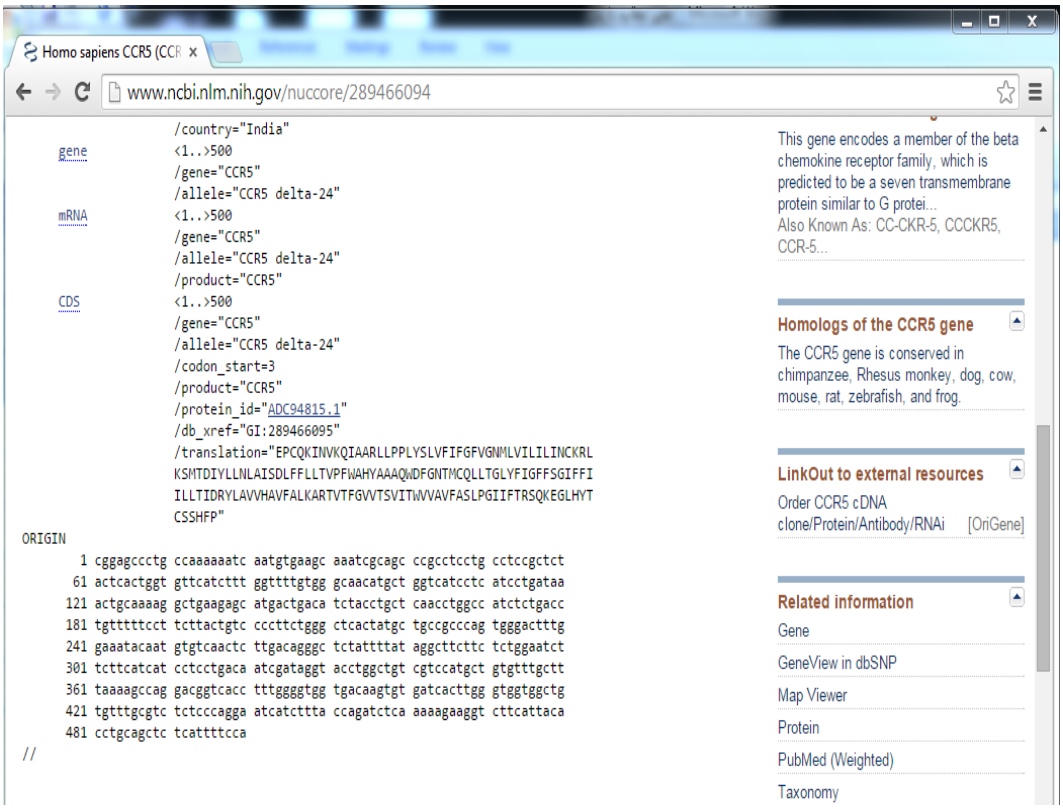
الجانب التطبيقي والتجريبي

3-1 المقدمة[8]

ان المشاهدات التي استخدمت في البحث تم الحصول عليها من موقع المركز الوطني لمعلومات التكنولوجيا الحيوية National Center for Biotechnology Information (NCBI) وهي عبارة عن متسلسلة من القواعد النيتروجينية في متسلسلة الحامض النووي DNA تمثل الجين CCR5 delta-24 allele الذي يؤثر في مقاومة الجسم لفيروس الايدز HIV وكما مبين في الشكل رقم (1) والشكل رقم (2)



الشكل (1)



الشكل (2)

ان هذه المتسلسلة هي عبارة عن متسلسلة من الاحرف (a , t , c . g) والتي يبلغ عددها (500) قاعدة نيتروجينية ، هذه القواعد النتروجينية تم تحويل الرموز الحرفية لها الى ارقام كالاتي :-

A ≡ 1 , t ≡ 2 , c ≡ 3 , g ≡ 4

ان القواعد النيتروجينية في متسلسلة الحامض النووي DNA تعتمد على مواقعها في المتسلسلة المطلوب التنبؤ بمتسلسلة الحالات من خلال متسلسلة الرموز الرقمية ومقارنتها بمتسلسلة الحالة الحقيقية .عنداستخدام نماذج ماركوف المخفية (HMMs) في بيانات المعلوماتية الحيوية يجب ان نحدد الأنموذج المستخدم وعند اختيار او تحديد الأنموذج يجب الاخذ بنظر الاعتبار الاتي :-

1- عدد الحالات والانتقالات

ان النموذج المستخدم يتكون من عدد محدد من الحالات ينبعث من كل حالة اربعة رموز بتوزيع احتمالي معين لكل رمز ، للأنموذج مجموعة من الانتقالات بين الحالات والتي تسمح بتغير الحالة بعد انبعاث الرموز .

الحالات (States) هي :- **N = (a , t , c , g)**

الاحتمالات الانتقالية (Transition Probability) هي :-

$$A=\begin{bmatrix} 0.320388 & 0.271845 & 0.23301 & 0.174757 \\ 0.092105 & 0.256579 & 0.328947 & 0.322368 \\ 0.270073 & 0.423358 & 0.240876 & 0.065693 \\ 0.185185 & 0.250000 & 0.277778 & 0.287037 \end{bmatrix}$$

الاحتمالات المنبعثة (Emission Probability) هي :-

$$B=\begin{bmatrix} 0.97 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.97 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.97 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.97 \end{bmatrix}$$

متسلسلة المشاهدات هي :-

O = (3441433324 3311111123 1124241143 1112343143 3343323324 3323343232
1323132442 4223123222 4422224244 4311312432 4423123323 1233241211
1324311114 4324114143 1241324131 2321332432 3113324433 1232324133
2422222332 2322132423 3332232444 3231321243 2433433314 2444132224
4111213112 4242311323 2241314443 2321222212 1443223223 2324411232
2322312312 3323324131 1234121442 1332443242 3423312432 4242224322
2111143314 4134423133 2224444244 2413114242 4123132244 4244244324
2422243423 2323331441 1231232221 3314123231 1111411442 3223122131
3324314323 2312222331)

2- الزمن Time

الزمن في متسلسلة الحامض النووي DNA يعني موقع المشاهدات على طول المتسلسلة .(2) وبتطبيق خوارزمية فيتربي Veterb على المعلومات (متسلسلة المشاهدات،الاحتمالات الانتقالية، الاحتمالات المنبعثة) باستخدام برنامج MATLAB تم التنبؤ بمتسلسلة الحالة الاتية :-

cggagccctg ccaaaaaatc aatgtgaagc aaatcgcagc ccgcctcctg cctccgctct
actcactggt gttcatcttt ggttttgtgg gcaacatgct ggtcatcctc atcctgataa
actgcaaaag gctgaagagc atgactgaca tctacctgct caacctggcc atctctgacc
tgtttttcct tcttactgtc cccttctggg ctcactatgc tgccgcccag tgggactttg
gaaatacaat gtgtcaactc ttgacagggc tctattttat aggcttcttc tctggaatct
tcttcatcat cctcctgaca atcgataggt acctggctgt cgtccatgct gtgtttgctt
taaaagccag gacggtcacc tttggggtgg tgacaagtgt gatcacttgg gtggtggctg
tgtttgcgtc tctcccagga atcatcttta ccagatctca aaaagaaggt cttcattaca
cctgcagctc tcattttcca

ان المتسلسلة المتنبئ بها هي مطابقة لمتسلسلة الحامض النووي للجين CCR5 delta-24 allele .

3-2 الاستنتاجات

1- ان نماذج ماركوف المخفية (HMMs) نماذج احصائية قوية للتنبؤ بمتسلسلة الحامض النووي DNA في المعلوماتية الحيوية Bioinformatics .
2- من خلال البحث تم التوصل الى ان المتسلسلة المتنبأ بها هي مطابقة لمتسلسلة الحامض النووي للجين CCR5 delta-24 allele.
3- كلما كانت معلمات النموذج مثلى او قريبة من الامثلية تكون متسلسلة الحالة المتنبأ بها صحيحة.

3-3 التوصيات

1- نوصي باختيار النموذج المناسب للبيانات وذلك من خلال تحديد الغرض من استخدام (HMMS) ، حيث هناك عدة نماذج لـ (HMMs) في المعلومانية الحيوية Bioinformatics .
2- نوصي باهتمام بحثي اكبر بالتعاون مع الاختصاصات ذات العلاقة في علم المعلوماتية الحيوية Bioinformatics .

المصادر

1- Birney,E.(2001)."Hidden Markov Models in Biological Sequence Analysis" ,IBM J.RES & DEV , Vol. 45 , No.3/4 ,P.449.
2- Valeria De Fonzo, Filippo Aluffi-Pentini and Valerio Parisi (2007) "Hidden Markov Models in Bioinformatics"Current Bioinformatics,Vol. 2,P. 49-61.
3- Smith , Kaleigh(2002) " Hidden Markov Models in Bioinformatics with Application to Gene Finding in Human DNA " January 17, 2002.
4- Rabiner.L.R and Juang.B.H. (1986)." An Introduction to Hidden Markov Models " IEEE ASSP MAGAZINE JANUARY 1986, 0740-7467/86/0100-0004$01.00@198I6E EE,P.4-16.
5- Stamp , M. (2012),"A Revealing Introduction to Hidden Markov Models", Associate Professor, Department of Computer Science, San Jose State University, Email: stamp@cs.sjsu.edu
6- Sean R Eddy(2004)" What is a hidden Markov model?" Nature Biotechnology Volume 22 Number 10 October 2004 .
7- Xiong , Jin (2006), " Essential Bioinformatics " Cambridge university press.
8- www.ncbi.nlm.nih.gov

……

……

……..