An Analytical Study of DNA Sequence

Basil Y. Thanoon
basyttb@yahoo.com
College of Computer Sciences and Mathematics
University of Mosul

Fatima M Hasan
Fatima.Zamzwm@yahoo.com
College of Computer Sciences and Mathematics
University of Mosul

Received on : 4/10/2010 **Accepted on : 10/11/2010**

ABSTRACT

This paper contains a general introduction to Bioinformatics, theirs goals, objectives and their applied fields. Is then shed light on the real views of DNA of humans, some processors use mathematical and statistical analysis of the traditional purpose of these observations and to identify some features and characteristics. Graphical analysis of DNA is carried out in two and three dimensions. The inter-connectedness among the sites of DNA is also analyzed as well as a spectral analysis. The attempt is also made to identify the order of such observations. Through this analytical study it is shown that DNA has a complex structure with interdependent with each other for long-term.

Keywords: Bioinformatics , spectral analysis , DNA

دراسة تحليلية لمتتابعة الحامض النووي الرايبي منقوص الأوكسجين

باسل يونس ذنون الخياط **فاطمة محمود حسن**

كلية علوم الحاسبات والرياضيات

جامعة الموصل

تاريخ استلام البحث: 2010/10/4 **تاريخ قبول البحث:2010/11/10**

الملخص

تتضمن هذه الورقة البحثية مقدمة عامة عن المعلوماتية الحيوية، وأهدافها ومضامينها ومجالاتها التطبيقية. كذلك يتم تسليط الأضواء على مشاهدات حقيقية للحامض النووي الرايبي منقوص الأوكسجين DNA للإنسان، وتستخدم بعض المعالجات الرياضية والإحصائية لغرض تحليل هذه المشاهدات والتعرف على بعض سماتها وخصائصها. إذ يتم إجراء تحليل بياني وتُرسم القواعد النيتروجينية للحامض النووي ببعدين وثلاثة أبعاد. كما يُدرس الترابط الداخلي بين مواقع الحامض النووي ويجرى تحليلا طيفيا لمشاهدات الحامض النووي. وتُجرى أيضا محاولة للتعرف على مَرتبة هذه المشاهدات. ومن خلال هذه الدراسة التحليلية يتبين بان مشاهدات الحامض النووي الرايبي منقوص الأوكسجين DNA للإنسان معقدة البنية التركيبية، ومترابطة مع بعضها البعض وعلى فترات طويلة الأمد.

الكلمات المفتاحية: المعلوماتيه الحيويه , الحامض النووي الرايبي منقوص الأوكسجين ,DNA, التحليل الطيفي

1.المعلوماتية الحيوية Bioinformatics :

تُعرّف المعلوماتية الحيوية على أنها استخدام الحاسوب لمعالجة المعلومات الحيوية. وهو علم ركيزته الأساسية قواعد بيانات المعلومات الحيوية بمكوناتها الرئيسية الجينات، والبروتينات، ويجمع عددا من العلوم الأخرى بهدف الاستفادة من هذه المعلومات كعلوم الرياضيات، والحاسوب ، والإحصاء، والطب، والكيمياء. ويمكن أن نلخص تعريف المعلوماتية الحيوية على أنها تطبيق التقانة الحاسوبية والمعلوماتية في إدارة المعلومات الحيوية. ومثال ذلك تحليل المعلومات الحيوية (الجينات والبروتينات) باستخدام الحاسوب والتقنيات الحاسوبية الحديثة [Shoemaker and Lin ,(2005)]. لقد أُستخدمت المعلوماتية الحيوية على نطاق واسع في الأبحاث التي تخص مشروع الجينوم البشري والذي حدّد السلسلة الجينيّة الكاملة للإنسان، و هو الثورة الجديدة التي تعد تغيير وجه الطب الذي نعرفه اليوم وعلاج أمراض لم يكن لها علاج من قبل. كما أن للمعلوماتية الحيوية دوراً كبيراً وفاعلا في اكتشاف عقاقير جديدة وفعّالة، حيث ساهمت بشكل كبير في إيجاد حلول لتحليل النتائج المختبرية المعقدة وكذلك استخدام الحاسوب لحفظ المعلومات واسترجاعها. كما أن البحث في المادة الوراثية للكائنات الحية يدخل ضمن نطاق المعلوماتية الحيوية، ويخدم هذا الفرع من العلم جميع العاملين في مجال الأبحاث العلمية الطبية والوراثية على حد سواء. كما يستخدم في عمليات تخزين البيانات وتحليل سلاسل الحامض النووي (DNA). ومن خلال المعلوماتية الحيوية يمكن أن نفهم بشكل أفضل الكيفية التي تنتظم فيها الجينات في سلسلة الحامض النووي (DNA). ويتوقع العلماء أن المستقبل القريب سيشهد تغيراً في الطريقة التي يعالج بها الأطباء مرضاهم، فبدلاً من إعطاء المريض مجموعة من الأدوية من خلال الفحص السريري، يمكن للطبيب معرفة الدواء الأفضل للمريض وتحديد حدة الاستجابة للدواء من خلال فحص المادة الوراثية للشخص ومقارنتها بالموجود في قواعد البيانات الجينية للأمراض ومن ثم التأكد إلى أي فئة ينتمي هذا المريض. وفي حالة الخطر يمكن تغيير الجرعة بدقة أو توجيه المريض لاستخدام دواء آخر وتغيير خطة العلاج بأكملها مما يجنب المريض مخاطر الأدوية قبل استخدامها (قاسم (2009)) .

تهدف المعلوماتية الحيوية إلى ثلاثة أهداف رئيسيّة وهي ((Shoemaker and Lin(2005):

1–تطوير تقنيات وبناء خوارزميات تساعد في تحصيل المعلومات من مجموعة ضخمة من البيانات.

2–تحليل وتفسير الأنماط المختلفة من البيانات التي تتضمن سلاسل الأحماض الأمينية والقطع والبنى البروتينيّة .

3–تطوير وتنفيذ أدوات تساعد على إدارة فعّالة للأنماط المختلفة من المعلومات.

تتضمّن المعلوماتية الحيوية المعالجة البارعة ، والتقصي ، والتنقيب عن بيانات Data Mining لمتتابعات DNA. إن تطوير التقنياتِ الخَزْن والتقصي عن متتابعات DNA أدّى إلى التقدّمِ الكبير جدا في الجانب التطبيقيِ في علوم الحاسوب، خصوصاً في مجالات خوارزميات تقصي سلسلة حروف String Searching Algorithms وتَعَلّم الماكنة Machine Learning ونظرية قاعدة البيانات Database Theory. إن التقصي عن متتابعات DNA يهتم بدراسة حدوثَ سلسلة حروف داخل سلسلة أكبر مِنْ الحروف، للتقصي عن متتابعات معيّنة مِن النكليوتيدات Nucleotides. وتستعمل سلاسل ماركوف لتشخيص الأشياء الشاذّة

Anomalies، وتصليح البيانات Repair Data، وتقييم سلامةَ البيانات Assess Data Integrity (الخياط (2010)).

من المجالات التطبيقية للمعلوماتية الحيوية ما يأتي ((Shoemaker and Lin (2005):

الطب الشخصي، العلاج الجيني، الطب الجزيئي، الطب الوقائي، العلاج باستخدام المورثات، تطوير الأدوية، تطبيقات الجينوم المايكروية، دراسات تغير المناخ، مصادر الطاقة البديلة، التقانة الحيوية، الممانعة المضادة الحيوية، تحسين المحاصيل، تحسين الجودة الغذائية، الطب البيطري.

2. الحامض النووي الرايبي منقوص الأوكسجين:

DNA هو مختصر Deoxyribonucleic Acid. يتألف جزيء الـ DNA من شريطين يلتفان حول بعضهما باتجاه عقارب الساعة، حول محور واحد، أحدهما يتجه إلى أعلى والآخر إلى أسفل، على هيئة سلم حلزوني مزدوج، كل شريط عبارة عن خيط من وحدات كيماوية تسمى النيوكلتيدات. والنيوكلتيدات تتكون من أربعة أصناف لا تختلف إلا في نوع القاعدة النيتروجينية، وهذه القواعد النيتروجينية هي:

الأدنين Adenin ، والكوانين Guanin ، والثايمين Thymine ، والسيتوسين Cytocine .

وتشكل هذه القواعد أزواجا ، فقاعدة "الأدنين" ترتبط دائما بـ "الثايمين"، بينما ترتبط "الكوانين" بـ "السيتوسين". وتتوزع القواعد بالترتيب على اللولب المزدوج، وتشكل القواعد كلمات وجملا وراثية تحفظ المعلومات الوراثية للكائن الحي من بداية الحياة إلى الممات، على هيئة جينات، وتتطابق كل مجموعة مؤلفة من ثلاثة أحرف مع حامض أميني واحد ، انظر الشكل (1) .

الشكل(1) : شريط DNA .

لقد كشفت الدراسات الحديثة أن للولب شريط DNA المزدوج، والذي يطلق علية أيضا جديلة، له خصائص مذهلة، لا سيما في العلاقة بين التركيب والوظيفة التي تؤكد أن التصميم الدقيق لهذا اللولب المزدوج المثير يشير بقوة إلى قدرة إبداع الخالق العظيم جل جلاله. فأنه إذا تم تمديد جديلة أل DNA الموجودة في أي خلية من خلايا الإنسان فسيبلغ طولها مترين. وإذا وضعت جميع جزيئات الحامض النووي للجسم البشري سوية من نهايات أطرافها فإنها قد تصل إلى الشمس وترتد أكثر من 600 مرة(Lindblad-Toh K, et al. (2005) ، انظر الشكل(2) .

الشكل(2) : تمديد جديلة DNA .

إن الحامض النووي الرايبي منقوص ألأوكسجين يمثل المادة الوراثية في نواة الخلية ، الحاوية لكل المعلومات الوراثية ، وقد أثبتت الدراسات تميز هذا الحامض ببعض الخصائص،منها ما يلي:

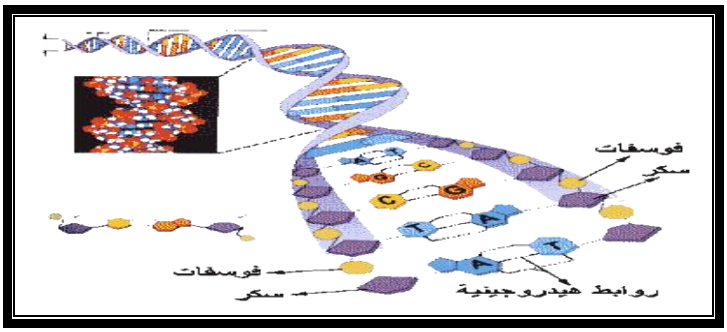
1 - إن كمية هذا الحامض ثابتة في جميع خلايا الأفراد، مهما كانت نوعية النسيج الذي يتكون منه العضو.

2 - إن لهذا الحامض قدرة على تكوين صورة طبق الأصل لنفسه، في أثناء الانقسام خلال المرحلة البينية، من خلال تفكك لولبه الشريطين المكونين له، بعد انفصام الروابط الهيدروجينية التي تربط بينهما، حيث يقوم كل شريط بعد ذلك بتكوين شريط مقابل، طبقاً للتجاذب النوعي للقواعد النيتروجينية، وينتج من هذا التكاثر الذاتي جزيئان متماثلان من هذا الحامض، مطابقان للجزيء الأصلي من حيث المكونات الأساسية.

3 - إن هذا الحامض يحتوي على جميع المعلومات الوراثية، توجد في ترتيبات وتعاقب ونوعية القواعد النيتروجينية على طول سلسلة الحامض، بمعنى أنه يمكن اعتبار هذه القواعد بمثابة لغة مكوَّنة من أربعة أحرف، يمكن استعمالها لتكوين كلمات مختلفة، حسب ترتيب القواعد الأربعة. إن هذه القواعد الأربعة تُرَمّز بالأحرف {A,G,C,T}، ويمكن تكوين جمل لها معنى من هذه الكلمات، ومن مجموعة هذه الجمل تتكون رسالة محددة ((Meyer (2009)).

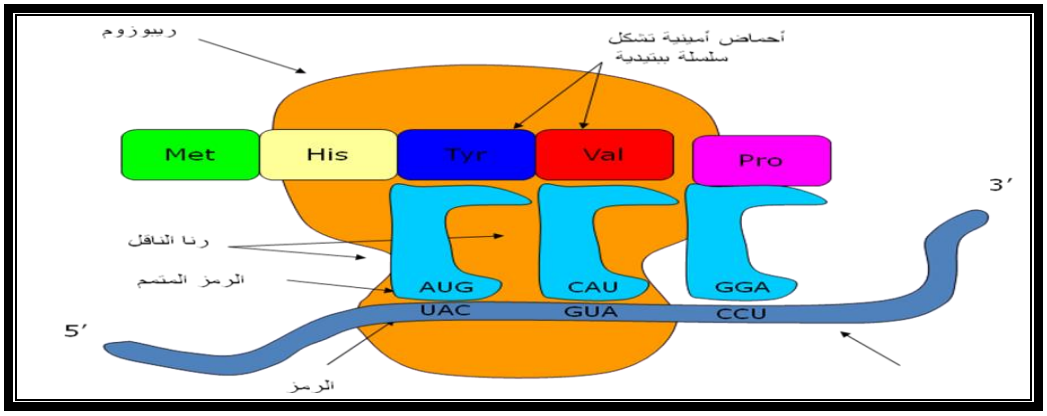
3. المتتابعة الزمانية للحامض النووي DNA:

إن أل DNA هو مادة وراثية موجودة في سائر الكائنات الحية وتحمل الصفات المورثة تواترا جيلا بعد جيل، وهو احد الخصائص الحيوية المهمة للكائنات الحية والتي تميز بعضها عن البعض الآخر. إن الحامض النووي DNA يتكون من شريطين ملتفين على بعضهما بحيث يشبهان السلم الملتوي، وأنه يتكون من أربعة أنواع من القواعد النيتروجينية هي, الأدنين (A) والثايمين (T) والسايتوزين (C) والكوانين (G), وتتكرر هذه القواعد ملايين أو مليارات المرات في جميع أجزاء الحامض النووي DNA ((Calvino, et al.(2007))، انظر الشكل(3).



الشكل(3): تركيب الحامض النووي (DNA).

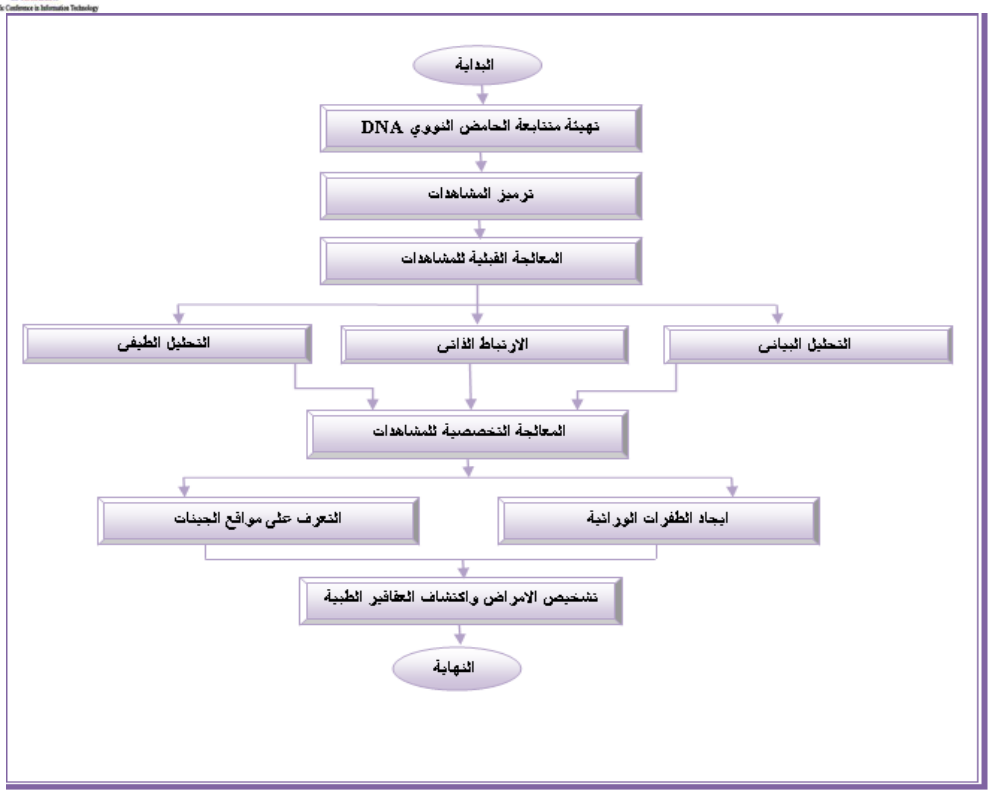
أن كل ثلاث قواعد نيتروجينية تعطي حامضا أمينيا واحدا ، كما أن عدد معين من الأحماض الامينية يؤدي إلى تكوين جين معين، وهذا الجين يكون مسئولا عن تكوين بروتين. وهذا البروتين بدوره له وظيفة محددة في معاني الحياة. يوجد أل DNA في نواة الخلية الحية كما يوجد عدد قليل منه في المايتوكوندريا (بيوت الطاقة) (قاري و جبر (2010))، انظر الشكل(4) .



الشكل (4): يمثل مقطع للمادة الوراثية DNA والتي تتضمنها الأحماض الامينية.

4. تحليل مشاهدات الحامض النووي:

إن المشاهدات التي تم استخدامها في هذا البحث هي عبارة عن سلسلة من القواعد النيتروجينية للحامض النووي DNA والموجودة في المايتوكندريا للإنسان الطبيعي . يبلغ حجم هذه المشاهدات 16571 قاعدة نيتروجينية بشكل متتابعة من الحروف الأربعة والتي هي G وC وT و A ، وهذه المعلومات متوفرة ضمن قاعدة بيانات (Data Base) في مراكز عالمية مختصة في الهندسة الوراثية ودراسة عمل الجينات والتي وفرت للباحثين عدداً من البيانات المتاحة لأجل الأبحاث والتطوير العلمي مثل [Gen Bank وNCBI] الشكل (5) يوضح المخطط البياني لمراحل عملية تحويل المشاهدات الحيوية لأستخدامها في التطبيقات الحاسوبية.



الشكل (5): المخطط البياني لمراحل عملية تحويل المشاهدات الحيوية لأستخدامها في التطبيقات الحاسوبية.

ملاحظة : وبالنظر الى ان المعالجة التخصصية تحتاج الى تفاصيل كثيرة ، لذا فان هذة الورقة البحثية سوف تخصص فقط على المعالجة القبلية .

4.1 التحليل البياني :

يمكن الاستعانة بالرسوم البيانية كأدوات مفيدة لأخذ فكرة أولية عن العلاقات الموجودة بين عناصر سلسلة القواعد النيتروجينية للحامض النووي. أن أول خطوة نحتاج إليها هي تحويل الرموز الحرفية الأربعة إلى أرقام . وقد تم إجراء هذا التحويل على النحو الأتي : $A \equiv 1$ و $T \equiv 2$ و $C \equiv 3$ و $G \equiv 4$.

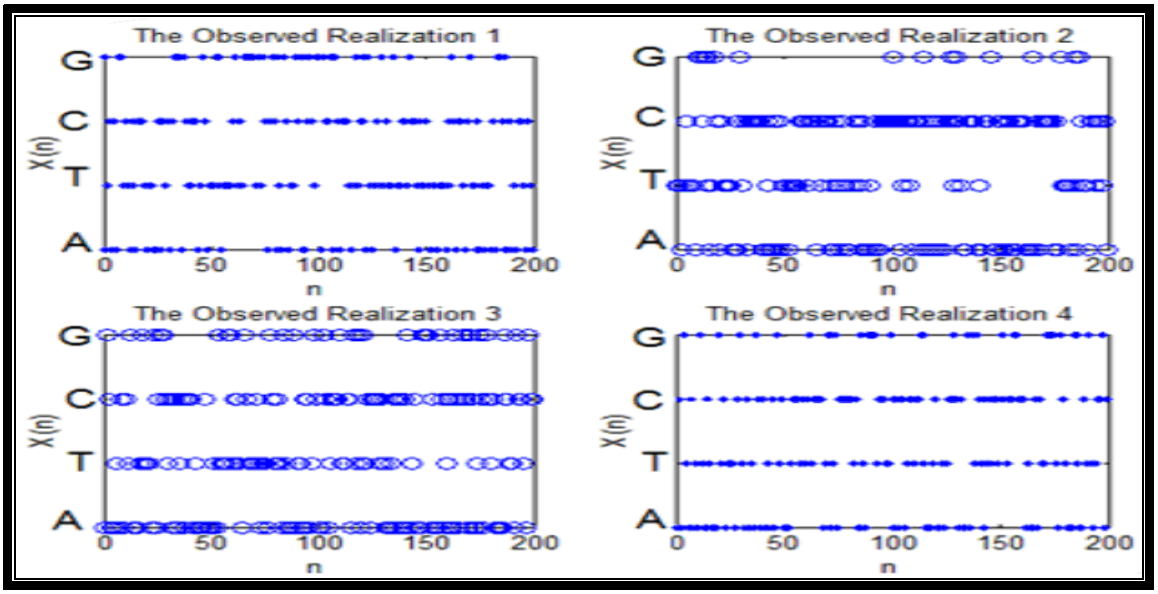
لأخذ فكره أولية عن الرسم الزمني لهذه المشاهدات فأن الشكل (6) يوضح أربعة مُتحققات مشاهدة من هذه المتتابعة طول كل منها *200* . وكما هو واضح فأن نمط انتشار النقاط في كل من هذه المتتابعات الأربعة يختلف عن الأخرى مما يشير إلى عدم وجود سلوك حتمي محدد لانتشار النقاط، وان هذه المشاهدات غير مُراوحة *Non-Stationary*، بمعنى أن خصائص القواعد النيتروجينية للحامض النووي تعتمد على مواقعها في سلسلة القواعد النيتروجينية للحامض النووي.

ولو فرضنا الآن إن المتغير العشوائي $X(n)$ *يمثل* القاعدة النيتروجينية في الموقع n من سلسلة القواعد النيتروجينية للحامض النووي، وبفرض أن القاعدة النيتروجينية في الموقع n، $X(n)$ ، هي دالة بدلالة القاعدة النيتروجينية التي تسبقها ب k من المواقع، $X(n-k)$ ، فيمكننا أن نفترض النموذج الرياضي الآتي بينهما:

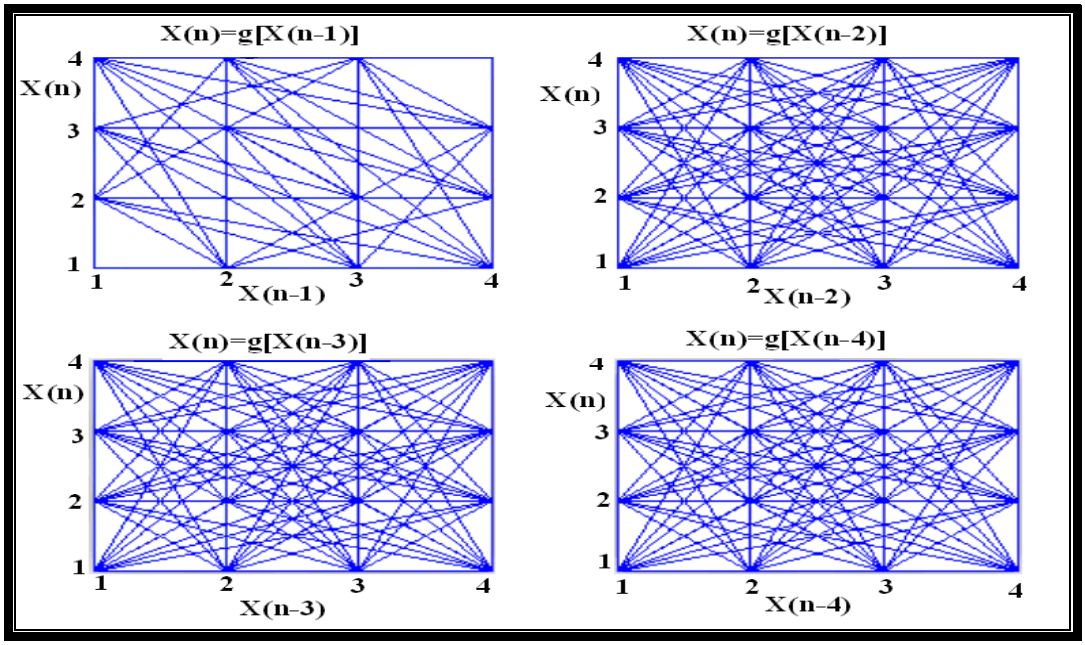
$$X(n) = g[X(n-k)];\ k = 1,2,3,... \qquad \ldots(1)$$

والشكل (*6a*) يوضح شكل الانتشار بين $X(n)$ و $X(n-k)$ للقيم k=1,2,3,4 *لمشاهدات* المتحققة *قيد* الدراسة. *أما* الشكل (*6b*) فيوضح شكل الانتشار بين $X(n)$ و $X(n-k)$ للقيم k=250,500,750,1000

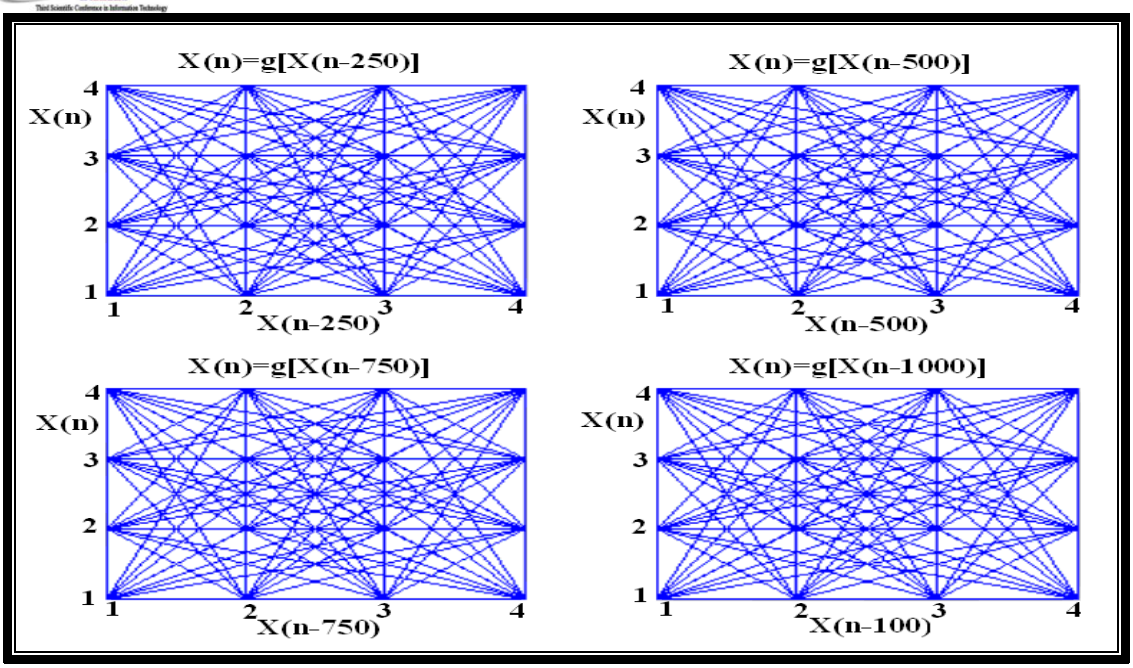
للمشاهدات نفسها. وكما هو واضح من الأشكال فهناك انتقالات بين جميع القواعد النيتروجينية الأربعة .من ناحية أخرى فمن الواضح أن العلاقة بين هذه القواعد تبدو ذات تركيب ذو طبيعة معقدة يختلف باختلاف قيمة k.



الشكل (*6*): أربعة مُتحققات مشاهدة من هذه المتتابعة طول كل منها *200* .



الشكل (*6a*): شكل الانتشار بين $X(n)$ و $X(n-k)$ للقيم k=1,2,3,4 *لمشاهدات* سلسلة القواعد النيتروجينية للحامض النووي.

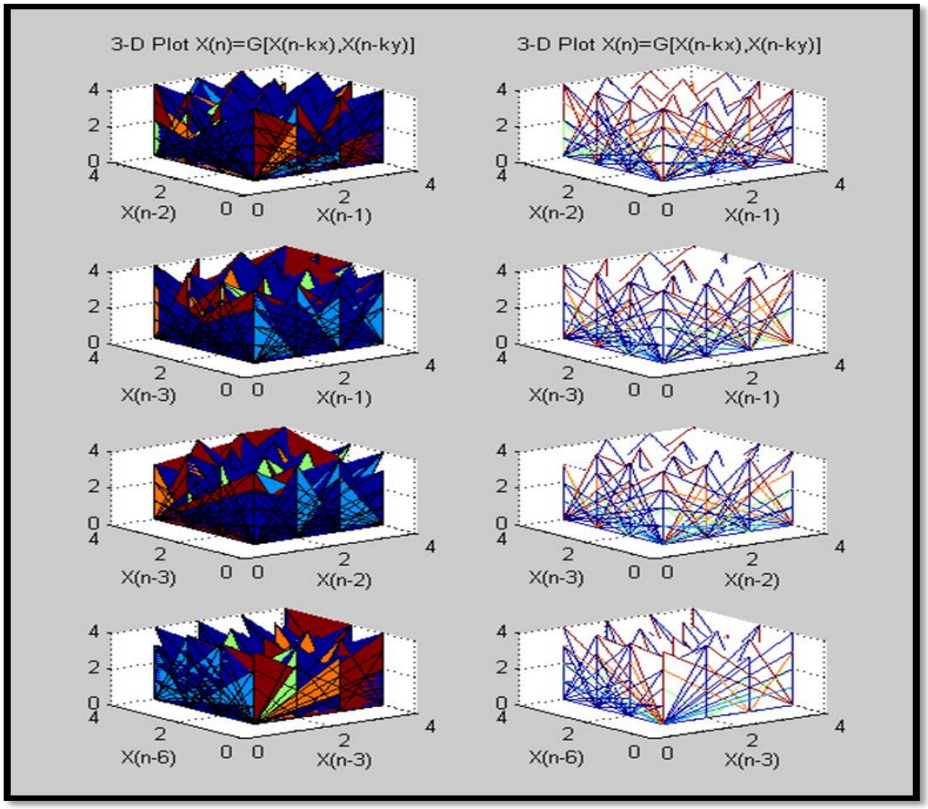


الشكل (*6b*): شكل الانتشار بين $X(n)$ و $X(n-k)$ للقيم k=250,500,750,1000 *لمشاهدات* سلسلة القواعد النيتروجينية للحامض النووي.

ويمكن تطوير الفكرة السابقة لمحاولة دراسة العلاقة بين القاعدة النيتروجينية في الموقع n وقاعدتين نيتروجينيتين في موقعين آخرين وذلك بفرض أن هذه العلاقة على النحو الآتي:

$$X(n) = G[X(n-kx), X(n-ky)];\ \ kx, ky = 1,2,3,... \qquad \ldots(2)$$

لقد تم إعداد برنامج خاص في MATLAB يقوم بتحويل المشاهدات (وهي ذات بُعد واحد One Dimensional) إلى منظومة ثلاثية البُعد Three Dimensional Array . والشكل (*7*) يبين النتائج التي حصلنا عليها لقيم مختارة من kx و ky ، وهذه النتائج تتضمن رسم الشبكة Mesh وكذلك السطح Surface الخاص والمقابل للقيم المختارة من kx و ky. إن العلاقة المعقدة بين مواقع القواعد النيتروجينية واضحة من خلال هذه الإشكال .



الشكل (7): الشبكة والسطح *الخاص والمقابل للقيم المختارة من* kx **و** ky .

4.2 الارتباط الذاتي Autocorrelation Function (ACF) :

تستخدم داله الارتباط الذاتي في تحليل مشاهدات المتسلسلات الزمنية. إذا كانت $\{x_1, x_2, ... x_T\}$ مشاهدات من متسلسلة زمنية معينة، فأن داله الارتباط الذاتي يمكن تقديرها على النحو الأتي (الخياط (2010)):

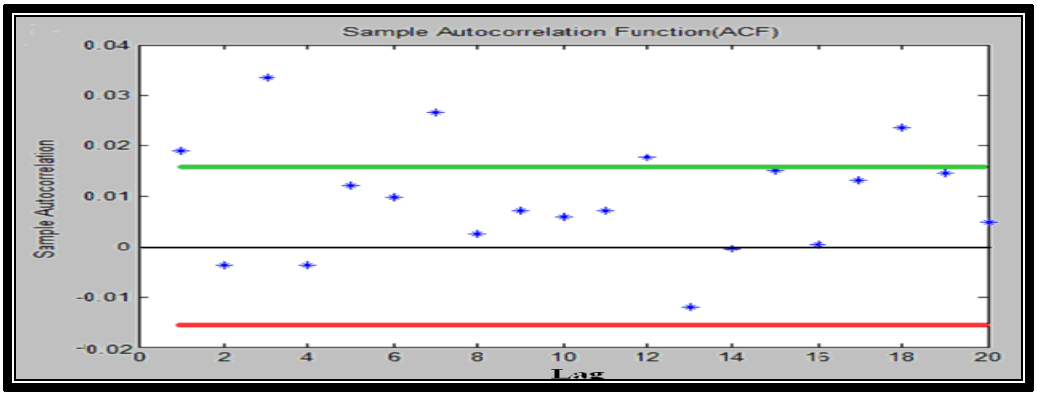
$$\hat{\rho}(k) = \frac{\sum_{t=1}^{T-k}(x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}; \quad k = 1,2,3,..., (T-1) \qquad \ldots(3)$$

إذ إن $\bar{x} = \sum_{t=1}^{T} x_t / T$ معدل المشاهدات.

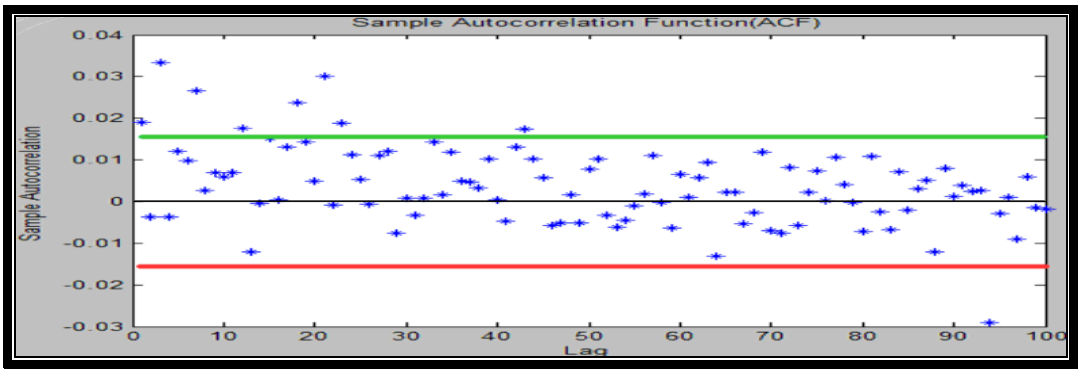
إن داله الارتباط الذاتي هي مقياس مجرد من الوحدات وتتراوح قيمه بين ± 1 , فإذا اقتربت قيمته من 1+ فأن ذلك يدل على أن هناك ترابطا طرديا قويا بين $x(n)$ و $x(n \pm k)$ ، وإذا اقتربت قيمته من 1- فأن ذلك يدل على أن هناك ترابطا عكسيا قويا بين $x(n)$ و $x(n \pm k)$. أما إذا كانت قيمة داله الارتباط الذاتي قريبة من الصفر، فان هذا دليل على عشوائية المتسلسلة. إن حدي الثقة Confidence Limits لدالة الارتباط الذاتي عند مستوى المعنوية 5% وتحت فرض عشوائية المتسلسلة الزمنية هي $\pm \frac{2}{\sqrt{T}}$.

أن دراسة الارتباط Correlation يعد احد الجوانب المهمة لدراسة العلاقة بين عناصر أي متتابعة، ويمكن انجاز ذلك من خلال ما يطلق عليه عادة اختبار العشوائية Test of Randomness. إن اختبار العشوائية يمكن انجازه بوساطة دالة الارتباط الذاتي المقدرة. وتجدر الإشارة، إلى أن العناصر الأكثر أهمية من دالة الارتباط الذاتي تكون لفترات الإبطاء الأولى، خاصة أول 20قيمة. إن أهمية دالة الارتباط الذاتي $\hat{\rho}_k$ من حيث الدلالة على الترابط الداخلي لقيم المشاهدات تتناسب مع اقتراب قيمة k من نقطة الأصل k=0. فكلما اقتربت قيمة k من الصفر كلما ازدادت أهمية $\hat{\rho}_k$. لذا فان أكثر قيم $\hat{\rho}_k$ أهمية هي $\hat{\rho}_1$ تليها من حيث الأهمية $\hat{\rho}_2$ ثم $\hat{\rho}_3$ على التوالي. وعند إجراء اختبار للعشوائية فيجب أن تكون قيم $\hat{\rho}_1$ و $\hat{\rho}_2$ و $\hat{\rho}_3$ تقع جميعها داخل الفترة $\pm\frac{2}{\sqrt{T}}$ لكي تكون فرضيتنا لعشوائية المشاهدات مقبولة بمستوى معنوية 5%. أما إذا وقعت في الأقل أي من $\hat{\rho}_1$ أو $\hat{\rho}_2$ أو $\hat{\rho}_3$ خارج الفترة $\pm\frac{2}{\sqrt{T}}$، فان فرضية عشوائية المشاهدات تُرفض حتى وان كانت جميع باقي قيم $\hat{\rho}_k$ تقع داخل فترة الثقة.

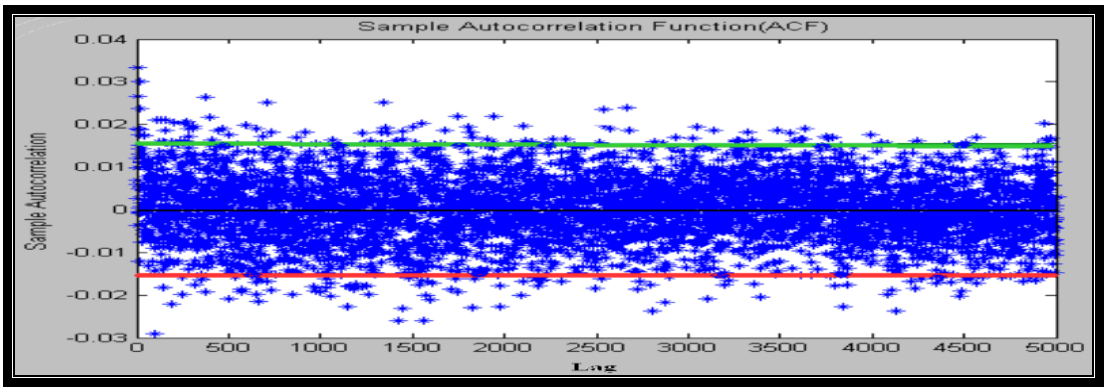
لقد تم تقدير و رسم دالة الارتباط الذاتي لمشاهدات القواعد النيتروجينية للحامض النووي وذلك بالاستعانة ب MATLAB. والشكل (8) يوضح رسم أول 20 وأول 100 وأول 5000 قيمة من هذه الدالة . فكما هو واضح فأن الترابط الداخلي موجود في فترات الإبطاء القريبة والمتوسطة والبعيدة. من ناحية أخرى فأن هذا الترابط ليس بالمعنوية الكبيرة وفي الوقت نفسه ليس بعدم المعنوية، بل هو أشبة ما يكون بترابط على حافات المعنوية، مما يؤكد العلاقة ذات الخصوصية الخاصة بين مواقع القواعد النيتروجينية في شريط الحامض النووي قيد الدراسة. إن الخطيين المتوازيين الذين تنتشر حولهما قيم دالة الارتباط الذاتي يمثلان حدي الثقة بمستوى معنوية 5%، فكما هو واضح فأن بعضا من النقاط تقع داخل حدي الثقة والبعض الأخر يكون خارجها وبالقرب من حافاتها , كذلك لا توجد قيم لدالة الارتباط الذاتي تزيد عن ± 0.1 . كل هذا يشير إلى أن هذه المشاهدات لا يمكن اعتبارها عشوائية وغير مترابطة مع بعضها البعض, من ناحية أخرى فلا يوجد ترابط قوي يمكن الاستناد علية لدراسة العلاقة بين مواقع هذه المشاهدات , وهذا يؤكد من جديد تعقيد البنية التركيبية لهذه المشاهدات.



الشكل (8a): دالة الارتباط الذاتي لمشاهدات القواعد النيتروجينية للحامض النووي لأول 20 قيمة.



الشكل (8*b*): دالة الارتباط الذاتي لمشاهدات القواعد النيتروجينية للحامض النووي لأول 100 قيمة.



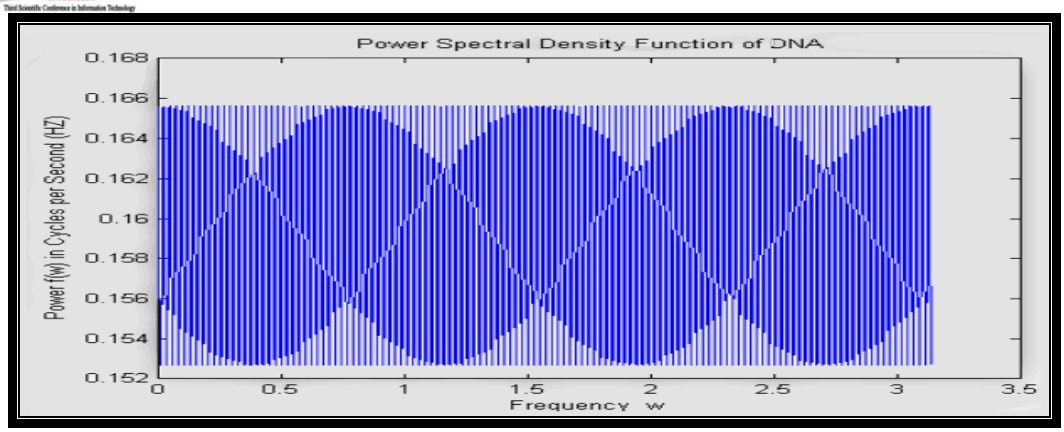
الشكل (8c): دالة الارتباط الذاتي لمشاهدات القواعد النيتروجينية للحامض النووي لأول 5000 قيمة.

4.3 التحليل الطيفي Spectral Analysis:

الشكل (10) يوضح مقدر لدالة كثافة الطيف المعيارية Normalized Power Spectral Density Function لمتتابعة الحامض النووي الرايبي منقوص الأوكسجين. وقد تم استخدام المقدر الآتي لهذا الغرض [Chatfield [1980]:

$$f(w)=\frac{1+2\sum \hat{\rho}_k \cos(kw_p)}{2\pi};\quad -2\pi \le w \le 2\pi,\ w_p=\frac{2\pi p}{T};\ p=1,2,...,\frac{T}{2}. \qquad ...(4)$$

وان M هي نقطة القطع Truncation Point، وقد اختيرت لكي تكون $M=[2\sqrt{T}]$ ، وان القوسين [] يعنيان إن يؤخذ الجزء الصحيح من المقدار المحصور بينهما.



الشكل (9) : مقدر دالة كثافة الطيف المعيارية لمتتابعة الحامض النووي الرايبي منقوص الأوكسجين.

نلاحظ من الرسم بأن الطاقة الموجودة في هذه المتتابعة تتوزع على مدى قصير يتراوح بين 0.152 و0.166 دورة لكل ثانية (هرتز). أن عدم وجود قمة وحيدة منفردة في هذه الدالة الطيفية يشير إلى عدم دورية هذه المتتابعة، أي أنها متتابعة غير دورية (Non Periodic). كما أن عدم ثبات مستوى الطاقة عند قيمة معينة يؤكد عدم عشوائية هذه المتتابعة. إن الشكل الذي حصلنا علية للدالة الطيفية لمتتابعة DNA يعد خاصة بهذه المتتابعة ويشير إلى أن هذه المتتابعة تتمتع ببنية تركيبة بالغة التعقيد خاصة بها.

4.4 تقدير المَرتبة Estimation of Order:

يعد تقدير مَرتبة سلسلة ماركوف من المسائل بالغة الأهمية في التطبيقات الواقعية، لما لها من علاقة بذاكرة السلسلة التي تتحكم ببعد النموذج. يقال للعملية التصادفية $\{X_t;\ t=0,1,2,\ldots\}$ بأنها سلسلة ماركوفية ذات مَرتبة محدودة m، أو ذات ذاكرة حجمها m ، حيث أن m عدد صحيح موجب اكبر من الصفر، إذا تحققت العلاقة الاحتمالية الآتية (Finesso (1991)):

$$P = (X_{t+1} = j \mid X_t = i_o, X_{t-1} = i_1, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots)$$
$$= P(X_{t+1} = j \mid X_t = i_o, X_{t-1} = i_1, \ldots, X_{t-m} = i_m) \quad , \qquad \ldots(5)$$

حيث أن $m>0$ ، $\forall i \in S$، وان $S=\{0,1,2,\ldots,N\}$ يمثل فضاء الحالة للسلسلة $\{X_t\}$.

ويمكن تعريف المَرتبة بأنها اقل عدد صحيح موجب ممكن لعدد الحالات السابقة المباشرة التي تعتمد عليها الحالة اللاحقة. فإذا ما اعتمد احتمال الحالة القادمة على m من الحالات السابقة مباشرةً، وكانت m اصغر عد صحيح موجب ممكن، عندئذ فان m ستمثل مَرتبة سلسلة ماركوف، وبحيث تحقق العلاقة (5). وقد تكون سلسلة ماركوف ذات مَرتبة صفرية (لا تمتلك ذاكرة)، أي أنها عبارة عن سلسلة من المتغيرات العشوائية (أو الحالات) المستقلة عن بعضها البعض.

لقد ظهرت في أواخر ستينيات وسبعينيات القرن العشرين معايير للمعلومات تعالج مسالة تقدير مَرتبة سلسلة ماركوف. ومن هذه المعايير المستخدمة معيار شانون للمعلوماتShannon Information Criterion، ومعيار خطأ التنبؤ النهائي Final Prediction Error ويرمز له اختصاراً FPE، ومعيار معلومات اكاكي Akaike's Information Criterion ويرمز له اختصاراً AIC، ومعيار معلومات ببز Bayesian Information Criterion ويرمز له اختصاراً BIC. وسوف يُعتمد في هذا البحث معيار AIC لتقدير مَرتبة متتابعة ألحامض

النووي منقوص الأوكسجين DNA. إن هذا المعيار هو مقياس لانحرافات النموذج عن النموذج الحقيقي. وان المَرتبة التي تقلل AIC تسمى بمقدّر أدنى AIC، Minimum AIC Estimator، والذي يرمز له اختصاراً MAICE. فلو كان MSE يمثل معدل مجموع مربعات البواقي، وكان T يمثل حجم العينة المستخدمة، و p يمثل عدد معلمات النموذج، فأن قيمة AIC العددية يمكن حسابها من المعادلة الآتية (Priestley(1981):

$$AIC(p)=T\ln(MSE)+2p \qquad \ldots(6)$$

إن الصيغة المكافئة لمعيار AIC في مجال تقدير مراتب سلاسل ماركوف تكون على النحو الأتي (الكسو (2005)):

$$R(k) = {}_k\eta_L - 2(\text{degrees of freedom}) \qquad \ldots(7)$$

إذ إن :

$$\left.\begin{aligned} {}_k\eta_L &= {}_k\eta_{k+1} + \ldots + {}_{L-1}\eta_L \\ &= \nabla^2 k_{k+2} + \ldots + \nabla^2 m_{L+1} \\ &= \nabla k_{L+1} + \ldots + \nabla k_{k+1} \qquad (-1 \le k < L) \\ &= -2\log \lambda_{k,\,k+1} - 2\log \lambda_{k+1,\,k+2} - 2\ldots - 2\log \lambda_{L-1,\,L} \end{aligned}\right\} \qquad \ldots(8)$$

وأن ∇ تمثل عامل الفرق Difference Operator. أما عدد درجات الحرية Degrees of Freedom *فيمكن* حسابها من الصيغة الآتية (Tong ,1975):

$$\begin{aligned} \text{Degrees of freedom} &= \nabla N^{L+1} - \nabla N^{k+1} \\ &\equiv N^{L+1} - N^{L} - N^{k+1} + N^{k} \end{aligned} \qquad \ldots(9)$$

حيث أن N تمثل عدد الحالات الممكنة للسلسلة في فضاء العينة .

من المعروف جيدا أن معيار AIC يجهز بمقدر غير متسق Not Consistent Estimate للمَرتبة الحقيقية للنموذج ، خاصة عندما يكون حجم المشاهدات كبير. وقد اقترح الباحثان Brockwell and Davis (1993) تصحيحا لمعيار AIC وأوصيا باستخدامه عندما يكون حجم المشاهدات كبير . إن معيار AIC المصحح يرمز له AIC_C ويحسب من العلاقة الآتية ((Brockwell and Davis (1991) :

$$AIC_C = AIC + \frac{2(k+1)(k+2)}{(n-k-2)} \qquad \ldots(10)$$

لقد تم تطبيق معياري AIC و AIC_C المعدلة على مشاهدات الحامض النووي الرايبي منقوص الأوكسجين وللرتب k=0,1,2، والنتائج مبينة في الجدول الأتي. ونشير إلى أن تطبيق هذه الطريقة على رتب أعلى يتطلب تكوين مصفوفات كبيرة جدا. فمثلا عندما تكون k=3 فإننا سوف نحتاج إلى تكوين مصفوفة ذات مَرتبة 256*256 . أما عندما تكون k=4 فأنة سوف يتطلب تكوين مصفوفة ذات مَرتبة 1024* 1024.

الجدول(1): تقدير مَرتبة متتابعة ألحامض النووي منقوص الأوكسجين .

k	Degrees of freedom	R(k)	AIC_C
0	**45**	**351.2360**	**351.2362**
1	**180**	**147.1142**	**147.1149**
2	**144**	**53.2684**	**53.2699**

نلاحظ أن هناك هبوط قوي في قيم AIC و AIC_C مما يؤكد الاعتمادية بين مواقع مَرتبة متتابعة ألحامض النووي منقوص الأوكسجين أل DNA . مع ذلك فنلاحظ بان هذه القيم مستمرة بالهبوط مما يشير إلى أن المَرتبة الحقيقية لهذه المتتابعة هي أكثر من2، إلا أن الصعوبات في تكوين المصفوفات لم تمكننا من الوصول إلى المَرتبة الحقيقية. ولعل استخدام التقنيات الذكائية يكون احد الوسائل التي توصل إلى المَرتبة الحقيقية، ولكن هذا خارج مجال بحثنا.

5. الاستنتاجات والتوصيات:

تضمنت هذه الورقة البحثية مقدمة عامة عن المعلوماتية الحيوية، وأهدافها ومضامينها ومجالاتها التطبيقية. ثم تم تسليط الأضواء على مشاهدات حقيقية للحامض النووي الرايبي منقوص الأوكسجين DNA للإنسان، واُستخدمت بعض المعالجات الرياضية والإحصائية لغرض تحليل هذه المشاهدات والتعرف على بعض سماتها وخصائصها. والاستنتاج الواضح بعد كل التحليلات التي أجريت في هذه الورقة البحثية هو أن هذه المتتابعة تتمتع بخصائص خاصة قل نظيرها: فهي متتابعة معقدة البنية التركيبية، ومترابطة مع بعضها البعض وحتى على الفترات البعيدة المدى. إن الأهمية الكبيرة للمتتابعات الزمانية لسلاسل الحامض النووي الرايبي منقوص الأوكسجين في عالمنا المعاصر من جهة، وهذا التركيب البنيوي المعقد لهذه المتتابعات من جهة أخرى، يدعو إلى التوصية بتوجّه بحثي اكبر، وبالتعاون مع التخصصات ذوات العلاقة، باتجاه لمتتابعات الزمانية لسلاسل الحامض النووي الرايبي منقوص الأوكسجين، واستخداما التقنيات والأدوات العلمية الحديثة، بما في ذلك التقنيات الذكائية أو التقنيات الحاسوبية الأخرى، لدراسة هذه المتتابعة بموضوعية اكبر وعمق أكثر. كما أن الأهمية البالغة للمعلوماتية الحيوية، وكما سلطنا الضوء على بعض جوانبها في هذه الورقة البحثية، تدعونا للتوصية بإنشاء مركز تخصصي في جامعة الموصل للمعلوماتية الحيوية يتضمن متخصصين بتخصصات علوم الحياة والطب وطب الأسنان والطب البيطري والزراعة، إضافة إلى تخصصات الحاسوب والرياضيات والإحصاء.

المصادر

[1]. الخياط ، باسل يونس ذنون (2010). "**النمذجة الماركوفية مع تطبيقات عملية**"، دار الكتب للطباعة والنشر، الموصل.

[2]. الكسو ،ابتهاج عبد الحميد (2005)."**استخدام الشبكات العصبية في تقدير رتب سلاسل ماركوف مع التطبيق على سلسلة جبل بطمة في محافظة الموصل** "،أطروحة دكتوراه غير منشورة ،كلية علوم الحاسبات والرياضيات ، جامعة الموصل.

[3]. قاري، سمير بن حسن محمد و جبر، جميل فوزي جميل (2010). " **مدخل والى الوراثة البشرية**" دار الفكر للطباعة والنشر , مكة المكرمة .

[4]. قاسم، عمر صابر (2009)." **تطبيق التقنيات الذكائية في المعلوماتية الحيوية** "، أطروحة دكتوراه غير منشورة، كلية علوم الحاسبات والرياضيات، جامعة الموصل.

[5]. Brockwell,R.J. and Davis, R.A. (1991). "Time Series: Theory and Methods",Springer, New York.

[6]. Calvino, M., Gomez, N. and Mingo, L.F.,(2007), "DNA Simulation of Genetic Algorithms: Fitness Computation", International Journal ,Information Theories & Applications, Vol.14.

[7]. Chatfield, C (1980):"The Analysis of Time Series: An Introduction", Chapman and Hall Ltd, London.

[8]. Finesso, Lorenzo. ,(1991): "Consistent estimation of the order for markov and hidden markov chains", Ph. D. , Dissertation.

[9]. http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html

[10]. Lindblad-Toh K, et al. (2005). "Genome sequence, comparative analysis and haplotype structure of the domestic dog .".*Nature* . 19-803:(7069) 438.

[11]. Meyer ,S. C.(2009)."Signature in the Cell: DNA and the Evidence for Intelligent Design", HarperColline_ books, ISBN 978-0-06-189421-3, USP

[12]. Priestley, M. B. ,(1981): "Spectral analysis and time series", Academic Press, INC. ,(London) LTD.

[13]. Shoemaker, J.S. and Lin, S.M., (2005), "Methods of Microarray Data Analysis IV", Springer Science + Business Media, Inc.

[14]. Tong, H. ,(1975): "Determination of the order of a Markov chain by using Akaike's information criterion", J. Appl. Prob. 12, 488-497.