Genetic Pattern Recognition using Intelligent Techniques

Basil Younis Thanoon **Omar Saber Qasim**

College of Computer Science and Mathematics
University of Al Mosul, Mosul, Iraq

Email: basytbb@Yahoo.com **Email: omar.saber@uomosul.edu.iq**

Received 16/08/2010; Accepted 10/11/2010

ABSTRACT

In this research a light is shed on converting DNA series to amino acid being responsible for forming protein through intelligent techniques. Some comparisons have been made between particularly Artificial Neural Network, Fuzzy Logic and Genetic Algorithms for discovering the powerful and the week ones in particle way. The hybrid operation has been made between Artificial Neural Network and Fuzzy Logic t to get a hybrid technique in a new formula having robust results than the original ones.

Keywords: Pattern recognition; neural network; fuzzy logic.

تمييز الأنماط الوراثية باستخدام التقنيات الذكائية

باسل يونس* الخياط** ***عمر صابر قاسم

كلية علوم الحاسوب والرياضيات
جامعة الموصل، الموصل، العراق

تاريخ استلام البحث: 2010/8/16 **تاريخ قبول البحث:2010/11/10**

الملخص

في هذه الورقة البحثية يتم تسليط الضوء على تحويل سلاسل الحامض النووي (DNA) إلى أحماض أمينيه مسئولة عن تكوين البروتينات وذلك باستخدام التقنيات الذكائية. كما يتم إجراء مقارنة بين عدد من هذه التقنيات, تحديداً الشبكات العصبية الاصطناعية والمنطق المضبب والخوارزمية الجينية, وذلك من اجل التعرّف على نقاط الضعف والقوة في كل منها في أثناء التطبيق. وبعد ذلك يتم إجراء عملية تهجين بين الشبكات العصبية الاصطناعية والمنطق المضبب للحصول على تقنية مهجنة بصيغة جديدة وبصفات ونتائج أفضل من الأصول. ويتبيّن من نتائج التطبيق العملي انه يوجد تفوق كبير للتقنية المهجَّنة عن تلك التقنيات غير المهجّنة.

الكلمات المفتاحية: تمييز الأنماط ; الشبكات العصبية الاصطناعية ; المنطق المضبب.

1-المقدمة

لتمييز الأنماط استخدامات واسعة ومتعددة, ولاسيّما الأنماط المتعلقة بالإنسان. إذ إن استخدام تمييز الأنماط في المعلوماتية الحياتية يعد نظاماً مساعداً للطبيب عندما تكون البيانات المدخلة كبيرة ومعقدة ويراد منها تشخيص حالة مرضية معينة, أو قد تكون البيانات المدخلة هي دلالة أو أساس لتكوين وبناء أجزاء حياتية مهمة مثل بيانات الحامض النووي (DNA), التي تترتب بشكل ثلاثيات وتحوّل إلى الأحماض الامينية. إن إصطفاف الأحماض الامينية بترتيب متناسق يعطي الأجزاء البروتينية داخل الجسم, فضلاً عن انه يمكن من خلالها التعرّف على الجينات الموجودة ضمن سلسلة الحامض النووي (DNA), وذلك من خلال مقارنتها مع مجموعة من الجينات الموجودة في قاعدة البيانات الوراثية.

2- الأنماط Patterns:

يمكن أن نعرف الأنماط أو النمط على أنه أي شيء (Object) يمكن أن يعطى له اسم, في حين أن صنف النمط (Pattern Class) هو مجموعة من الأنماط التي تشترك بخصائص معينة تنشأ من نفس المصدر, وفي أثناء عملية التصنيف يتم ترتيب الأنماط وفقاً للأصناف التي تنتمي إليها وذلك من خلال بناء آلة (Machine) تسمى المصنف (Classifier) والتي تقوم بمهام عملية التصنيف [Joshi, et al.; 1997].

3- مرحلة تصنيف الأنماط The Patterns Classification Stage :

هي عملية تجميع الأشياء المتشابهة معاً، ويشترك جميع أعضاء المجموعة الواحدة أو القسم الواحد أو الصنف الواحد الناتج عن التصنيف في خاصية واحدة على الأقل لا يملكها أعضاء الأقسام أو الأصناف الأخرى. وقد يجري تحديد التشابه عن طريق الحاسوب أو عن طريق الإنسان بواسطة الإدراك البشري للخصائص المشتركة.

4- وصف البيانات الوراثية :

إن البيانات المستخدمة عبارة عن جينات أو سلسلة من الحامض النووي (DNA), وهذه المعلومات متوفرة ضمن قاعدة بيانات (Data Base) في مراكز عالمية مختصة في الهندسة الوراثية ودراسة عمل الجينات , مثل NCBI و GenBank و EMBL . لقد تم تشفير القواعد النتروجينية الأربعة التي بتتابعاتها المختلفة تتشكل سلسلة بيانات الحامض النووي وذلك ضمن إطار عالمي محدد من قبل مراكز عالمية مختصة مثل الإتحاد الدولي للكيمياء الصرفة والتطبيقية (International Union of Pure and Applied Chemistry) والذي يرمز له IUPAC اختصاراً.

5- تشفير سلاسل الحامض النووي (DNA):

إن السلاسل الزمانية للحامض النووي تتكون من أربعة رموز رئيسية هي (A,C, G, T), إذ إن كل ثلاثية متتالية من هذه الرموز تُتَرجم إلى حامض أميني وحيد. وبما أن عدد أنواع النيوكليوتيدات هو أربعة, لذا فإن عدد الثلاثيات المختلفة التي يمكن تكوينها هو (64). ولما كان عدد الحوامض الامينية هو (20) نوعاً,

لذلك فمن الواضح أنّ هناك حوامض أمينية تحددها أكثر من ثلاثية واحدة من رموز الحامض النووي (DNA)، كما في الجدول (1) الآتي:

الجدول (1) : يبين تشفير سلاسل الحامض النووي (DNA) إلى الأحماض الامينية.

	T	C	A	G
T	**TTT** - Phe (F) **TTC** - Phe (F) **TTA** - Leu (L) **TTG** - Leu (L)	**TCT** - Ser (S) **TCC** - Ser (S) **TCA** - Ser (S) **TCG** - Ser (S)	**TAT** - Tyr (Y) **TAC** - Tyr (Y) **TAA** - **STOP** **TAG** - **STOP**	**TGT** - Cys (C) **TGC**- Cys (C) **TGA** - **STOP** **TGG** - Trp (W)
C	**CTT** - Leu (L) **CTC** - Leu (L) **CTA** - Leu (L) **CTG** - Leu (L)	**CCT** - Pro (P) **CCC** - Pro (P) **CCA** - Pro (P) **CCG** - Pro (P)	**CAT** - His (H) **CAC** - His (H) **CAA** - Gln (Q) **CAG** - Gln (Q)	**CGT** - Arg (R) **CGC** - Arg (R) **CGA** - Arg (R) **CGG** - Arg (R)
A	**ATT** - Ile (I) **ATC** - Ile (I) **ATA** - Ile (I) **ATG** - Met (M) **START**	**ACT** - Thr (T) **ACC** - Thr (T) **ACA** - Thr (T) **ACG** - Thr (T)	**AAT** - Asn (N) **AAC** - Asn (N) **AAA** - Lys (K) **AAG** - Lys (K)	**AGT** - Ser (S) **AGC** - Ser (S) **AGA** - Arg (R) **AGG** - Arg (R)
G	**GTT** -Val (V) **GTC** -Val (V) **GTA** -Val (V) **GTG** -Val (V)	**GCT** - Ala (A) **GCC** - Ala (A) **GCA** - Ala (A) **GCG** - Ala (A)	**GAT** - Asp (D) **GAC** - Asp (D) **GAA** - Glu (E) **GAG** - Glu (E)	**GGT** - Gly (G) **GGC** - Gly (G) **GGA** - Gly (G) **GGG** - Gly (G)

6-استخدام التقنيات الذكائية للحصول على الأحماض الامينية:

تعد البروتينات من الجزيئات الكبيرة والمعقدة والمكوّنة من سلسلة طويلة من القطع التي تسمى بالأحماض الأمينية, كما تعد من أهم المواد التي تقوم بتشكيل صفات الكائن الحي, فضلاً إلى مسؤوليتها عن بناء الخلية واستمرارها في العمل. كما أن الاختلاف في طريقة تتابع الأحماض الأمينية هو الذي يكوّن البروتينات المختلفة. ويمكن توضيح الخطوات العامة لتحويل سلسلة الحامض النووي (DNA) إلى الأحماض الامينية المسئولة عن تكوين البروتينات من خلال المخطط الانسيابي الآتي :

الشكل (1) : المخطط الانسيابي للخطوات العامة لتحويل سلاسل الحامض النووي (DNA).

لقد تم تدريب كل من تقنية الشبكات العصبية (Neural Networks), والمنطق المضبب (Fuzzy Logic), والخوارزمية الجينية (Genetic Algorithm) فضلاً عن تقنية مهجنة مابين المنطق المضبب والشبكات العصبية (Neuro Fuzzy) في الحصول على الأحماض الامينية (Amino Acid). كما تم أخذ جميع الاحتمالات الثلاثية المتكونة من الرموز الرئيسية الأربعة (A,C,G,T) للحامض النووي (DNA), إذ أن مجموع جميع الاحتمالات هي (64) حيث أن: $4^3 = 64$.

7- تطبيق الشبكات العصبية على بيانات الحامض النووي (DNA):

تم اعتماد شبكة الانتشار الخلفي للخطأ (BP) في تمثيل بيانات سلسلة الحامض النووي (DNA), ولكون هذه الخوارزمية تحتاج إلى مدخلات عددية, في حين أن سلسلة الحامض النووي (DNA) عبارة عن متغيرات حرفية, لذا فقد تم تشفير هذه المتغيرات الحرفية إلى أعداد صحيحة وذلك لكي نستطيع التعامل معها بوصفها مدخلات تتناسب مع الشبكة العصبية.

8- تهيئة أفضل المعلمات للشبكة العصبية (BP) :

تتكون معمارية الشبكة العصبية الاصطناعية (BP) من مجموعة من المعلمات التي تساعد في تطوير تدريب خوارزميتها في أثناء تعديل الأوزان, ولكون معظم هذه المعلمات عبارة عن صيغ أو طرق رياضية تساعد في تقليل الخطأ الناتج في أثناء عملية التدريب, لذا فقد تمت مقارنة المعلمات في أثناء التطبيق لإيجاد أفضلها واعتبارها الصيغ الأساسية التي تبنى عليها الشبكة عند مقارنتها مع التقنيات المختلفة الأخرى. ومن هذه المعلمات :

1-عدد الطبقات المخفية (Hidden Layers) المستخدمة في معمارية الشبكة .

2-عدد العقد (Nodes) في الطبقة المخفية.

3- نوع دوال التنشيط (Activation Functions) المستخدمة في الطبقة المخفية وطبقة الإخراج.

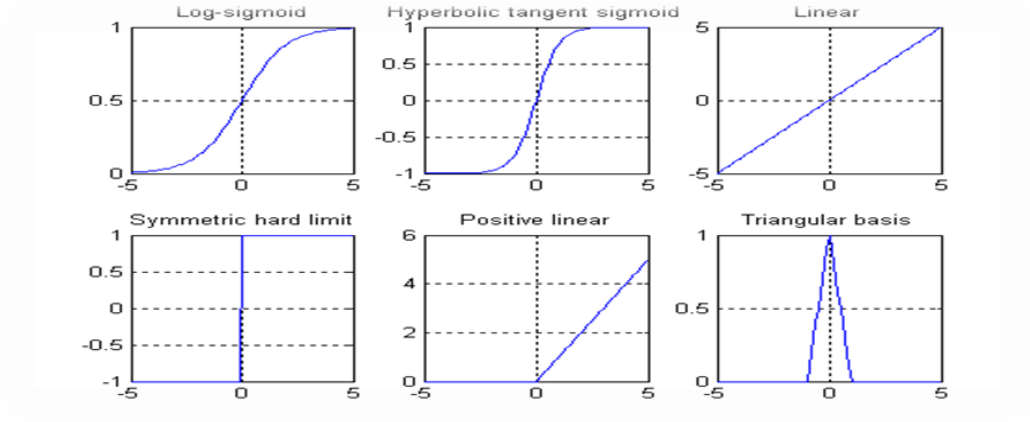
4- نوع الخوارزمية المستخدمة في تدريب الخطأ (Error Training).

9- تحديد عدد العقد ودوال التنشيط المستخدمة في الطبقة المخفية :

إن عدد العقد في الطبقة المخفية له تأثير كبير في عملية معالجة وتدريب الشبكة العصبية (BP). لقد تم مقارنة أعداد مختلفة من العقد المستخدمة في الطبقة المخفية أثناء التدريب لإيجاد العدد الأفضل المستخدم في تمثيل معمارية الشبكة, حيث أن المقارنة اعتمدت في ذلك على مقياس الكفاءة (Efficiency). حيث أن كفاءة أي طريقة أو تقنية ذكائية تتناسب عكسياً مع كل من الزمن ومعدل مربع الخطأ للنتائج, لذا يمكن تعريف الكفاءة على النحو الآتي:

$$EFF = 1/(MSE * Time \quad \ldots(1)$$

إن أنواع مختلفة من دوال التنشيط تم مقارنتها في الطبقة المخفية لإيجاد أفضل النتائج للشبكة العصبية, وقد تبين أثناء التطبيق أن دالة التنشيط (Positive linear) هي الأفضل ضمن بيانات الحامض النووي (DNA), وذلك وفق مقياس الكفاءة. كما يمكن توضيح مدى دوال التنشيط المستخدمة في اختبار الشبكة من خلال الشكل (2) الآتي:



الشكل (2) : يوضح دوال التنشيط المستخدمة في اختبار الشبكة العصبية.

إن كل عقدة في الطبقة المخفية وطبقة الإخراج تحتوي على دوال التنشيط, ونوع هذه الدوال قد يختلف من طبقة إلى أخرى. ولأن المدى في دوال التنشيط المختلفة يحدد إخراج الشبكة, لذا تم تثبيت الدالة الخطية (Linear Function) في طبقة الإخراج لان مداها يسمح بمرور جميع الأعداد الصحيحة والتي من ضمنها نواتج الشبكة.

الجدول (2) : يوضح اختيار أفضل دالة تنشيط وأفضل عدد للعقد في الطبقة المخفية.

نوع دالة التنشيط ***Type of Activation Function***	*عدد العقد في الطبقة المخفية*	*معدل مربع الخطأ* ***MSE***	*الزمن بالثانية* ***Time***	*مقياس الكفاءة* ***EFF***
Log-sigmoid	38	19.016	0.686	0.0767
Hyperbolic tangent sigmoid	4	25.530	0.739	0.0530
Linear	25	39.483	0.565	0.0448

0.0444	0.532	42.343	45	Symmetric hard limit
0.1234	0.776	10.445	15	Positive linear
0.0762	0.738	17.792	5	Triangular basis

من الجدول (2) يتضح لنا بأن أفضل مقياس للكفاءة (EFF) يكون عند الدالة الخطية الموجبة (Positive linear), وان أفضل اختيار لعدد العقد في الطبقة المخفية هو (15) عقدة. إذ أننا قمنا باختبار تمثيل عدد العقد في الشبكة ابتداءً من عقدة واحدة وانتهاءً عند خمسين عقدة لكل دالة في الطبقة المخفية, وتبين لنا أن أفضل النتائج المستحصلة من خلال مقياس الكفاءة هي القيم المسجلة في الجدول (2).

10- أنواع خوارزميات تدريب الخطأ في شبكة الانتشار الخلفي (BP) :

لقد استخدمت العديد من خوارزميات تعديل أوزان الشبكة العصبية بالاتجاه العكسي مثل خوارزمية (Levenberg-Marquardt), وخوارزمية التدرج المترافق (Conjugate Gradient) التي تتضمن (Ribiere Polak , Fletcher-Reeves , Powell-Beale restarts) وكذلك خوارزمية (Gradient descent) فضلاً عن خوارزمية (BFGS) و (quasi-Newton). وتبين أثناء التطبيق العملي أن هذه الخوارزميات تختلف من حيث سرعة التنفيذ ومقدار خطأ التدريب, فضلاً عن مقياس الكفاءة المعتمد على عاملي الوقت وكمية الخزن. إذ إن خوارزميات تعديل الأوزان تم تطبيقها في شبكة الانتشار الخلفي للخطأ (BP) على سلسلة الحامض النووي (DNA), ضمن برنامج الماتلاب (MATLAB). والجدول (3) الآتي يبين المقارنة بين الخوارزميات المذكورة آنفاً:

الجدول (3) : مقارنة بين طرق تعديل أوزان الشبكة العصبية الاصطناعية.

خوارزمية التدريب *Training Algorithm*	معدل مربع الخطأ *MSE*	الزمن بالثانية *Time*	مقياس الكفاءة *EFF*
Levenberg-Marquardt	10.445	0.982	0.0975
Powell-Beale restarts	32.613	0.775	0.0396
Fletcher-Reeves	15.779	1.220	0.0519
Polak-Ribiere	41.664	0.719	0.0334
Gradient descent	657.91	0.550	0.0028
BFGS quasi-Newton	32.467	0.983	0.0313

يتضح لنا من الجدول (3) أن أفضل اختيار لخوارزمية التدريب المستخدمة في تعديل أوزان الشبكة العصبية الاصطناعية هي خوارزمية (Levenberg-Marquardt) التي يكون عندها مقياس الكفاءة هو الأفضل.
كما يتضح لنا من الجدولين (2) و (3) السابقين أن أفضل معمارية للشبكة العصبية الاصطناعية (BP), بالنسبة لبيانات الحامض النووي (DNA), هي التي تعطي أفضل مقياس للكفاءة من خلال المعلمات الآتية :

الجدول (4) : يوضح أفضل معلمات شبكة (BP) المستخدمة في توليد الأحماض الامينية.

نوع الشبكة العصبية الاصطناعية	شبكة الانتشار الخلفي للخطأ (BP)
عدد العقد في الطبقة المخفية	15
نوع الدالة في الطبقة المخفية	Positive linear
نوع الدالة في طبقة الإخراج	Linear
نوع خوارزمية التدريب	Levenberg-Marquardt

11- تطبيق الخوارزمية الجينية على بيانات الحامض النووي (DNA):

تعد الخوارزمية الجينية من طرق البحث الكفوءة التي تعتمد على مبادئ الانتخاب الطبيعي وعلم الوراثة, وكان الهدف الأساس من تصميمها هو بناء العديد من الخوارزميات والبرمجيات التي تقوم بتحسين الحل. وقد تم تطبيق الخوارزمية الجينية على مصفوفة تتكون من (64) صفاً و(3) أعمدة تمثل جميع الحالات المختلفة من رموز الحامض النووي (DNA), وهي (A,C,G,T), وذلك لتكوين الحوامض الامينية (Amino Acid). وقد أثبتت كفاءة عالية في تصنيف إدخالات التدريب, لأجل الوصول لأفضل الحلول, لكن المشكلة في الخوارزميات الجينية أنها تفتقد إلى عملية المحاكاة (Simulation), وبمعنى آخر, أن عملية الاختبار (Test) لقيم جديدة لايمكن إجرائها. وبهذا لايمكن الحصول فيها على نموذج رياضي قادر على توقع سلسلة الحوامض الامينية بالنسبة لسلسلة الإدخالات إلا من خلال معالجة كاملة بكل خطوات الخوارزمية الجينية. ولهذا لايمكن اعتمادها في تحليل بيانات الحامض النووي (DNA) لأنها ستستغرق وقتا طويلا في معالجة كل إدخال لتلك البيانات. والجدول (5) يوضح المقارنة في مقياس الكفاءة (EFF), بين شبكة الانتشار الخلفي للخطأ (BP) والخوارزمية الجينية (GA).

الجدول (5) : مقارنة بين شبكة (BP) والخوارزمية الجينية في الحصول على الأحماض الامينية.

نوع التقنية الذكائية *(Type of Technique)*	معدل مربع الخطأ *(MSE)*	الزمن بالثانية *(Time)*	مقياس الكفاءة *(EFF)*
الخوارزمية الجينية (GA)	7.501 e-5	8.661	1.539 e+3
الشبكة العصبية الاصطناعية (BP)	10.445	0.982	0.0975

من الجدول (5) السابق يتضح لنا أن نتائج الخوارزمية الجينية أفضل بكثير من خلال مقياس الكفاءة من نتائج شبكة الانتشار الخلفي للخطأ (BP), إلا أن الخوارزمية الجينية لا تستطيع محاكاة أية سلسلة بتتابعات مختلفة وفقاً لحالات الإدخال. ولان الغاية الأساسية من تدريب التقنيات الذكائية هي عملية المحاكاة (Simulation), التي تحصل بعد التدريب على جميع الحالات الثلاثية لسلسلة الحامض النووي (DNA), ولهذا كان لابد من تطوير الشبكة العصبية (BP) أو تهجينها للحصول على نتائج جيدة من خلال مقياس

الكفاءة المعتمد على معدل مربع الخطأ والزمن المستغرق في تنفيذ البرنامج فضلاً عن كونها تقوم بعملية المحاكاة.

12- تطبيق المنطق المضبب على بيانات الحامض النووي (DNA):

تعد تقنية المنطق المضبب من التقنيات التي تتمتع بقدرة عالية على إيجاد الحلول للعديد من التطبيقات الحياتية التي تخضع لقواعد وأطر علمية متعارف عليها. كما يمكن الحصول من خلالها على استنتاجات محددة, ولو كانت بيانات الإدخال غامضة أو غير دقيقة, وذلك من خلال محاكاة قدرات الإنسان في اتخاذ القرار عن طريق تحويل الإدخال إلى مجموعة مضببة باستخدام دوال العضوية. فضلاً عن عملية الاستدلال المنطقي التي تعتمد على قواعد (IF THEN) وذلك من خلال شروط معينة تستند عليها بيانات الإدخال للحصول على الإخراج المطلوب.

تتضمن تقنية المنطق المضبب عدداً من المعلمات القابلة للتغيير كما في الشبكات العصبية الاصطناعية, وذلك من أجل التعامل مع التطبيقات المختلفة بشكل يتلاءم وطبيعة المسألة المعطاة. فمثلا بيانات الإدخال للحامض النووي (DNA), متكونة من مصفوفة يمكن أن نرمز لها بالحرف (P), وهي بسعة (64) صفاً و (3) أعمدة. كما يمكن التعبير عنها برمجيا بالشكل الآتي :

$$P = [p1; p2; p3] \quad , \quad p1 = P(:,1) \, , p2 = P(:,2) \, , \, p3 = P(:,3)$$

عند بناء نموذج مضبب لبيانات الحامض النووي, تظهر لدينا مشكلة في تحديد دالة العضوية المناسبة لكل إدخال ($p_i \, , i = 1,2,3$) وذلك من خلال إختيار أمثل المعلمات لتكوين شكل هذه الدالة, فضلاً عن تحديد نوعها. إذ يوجد عدد كبير من الاحتمالات في اختيار دوال العضوية يصل عددها إلى (n^3) من الاختيارات, إذ أن قيمة (n) تمثل عدد الدوال التي يمكن تمثيلها في الإختيارات الثلاثة ($p1, p2, p3$). والأهم من هذا كله هو اختيار معلمات هذه الدالة, فمثلا دالة كاوس (Gauss) تحتاج إلى اختيار قيمة ($c\sigma$,) التي تناسب طبيعة الإدخال للبيانات, وكذلك دالة شبه المنحرف (Trapezoidal) تحتاج لاختيار القيم (a,b,c,d) المناسبة للإدخال حين تمثيلها.

يتضح مما سبق أن مشكلة الاستدلال المضبب تكمن في تعديل معلمات دوال العضوية في العديد من التطبيقات التي لاتخضع لقوانين عامة, أو عندما يكون فيها تعقيد أثناء بناء القواعد (Rules). وقد تبين لنا أثناء التطبيق العملي على سلسلة الحامض النووي (DNA), أن المنطق المضبب لوحده لايعطي نتائج مقبولة أو حتى مرضية.

13- تهجين التقنيات الذكائية :

تم تطبيق كل من الشبكات العصبية (Neural Networks) والخوارزمية الجينية (Genetic Algorithm) والمنطق المضبب (Fuzzy Logic), على بيانات سلسلة الحامض النووي (DNA), واتضح لنا من خلال التطبيق عدد من نقاط الضعف والقوة في كل من هذه التقنيات. وسنركز في هذه الدراسة من خلال التقنيات المذكورة سابقاً على ثلاثة معايير أساسية, فضلاً عن جوانب مهمة ومشتركة في كل من هذه التقنيات وهي :

1- معيار معدل مربع الخطأ (MSE).

2- مقدار الوقت المستغرق في معالجة البيانات (Time).
3- إمكانية إجراء عملية المحاكاة (Simulation).

ومن خلال التركيز على هذه المعايير الثلاثة, يمكن بناء تقنية ذكائية مهجنة تجمع نقاط القوة الموجودة في التقنيات الذكائية وتعطي نتائج جيدة وفق مقياس الكفاءة, فضلاً عن إمكانية إجراء عملية المحاكاة للبيانات الجديدة. ولأجل إجراء تهجين بين التقنيات التي تم طرحها, فإن الجدول (6) يوضح مقارنة بين هذه التقنيات من خلال المعايير الثلاثة (دقة النتائج , الوقت , المحاكاة).

الجدول (6) : مقارنة بين الشبكات العصبية والخوارزمية الجينية والمنطق المضبب عند التطبيق.

نوع التقنية (Type of Technique)	*عملية المحاكاة (Simulation)*	*معدل مربع الخطأ (MSE)*	*وقت التنفيذ (Time)*
الشبكة العصبية (NN)	تتضمن	كبير	قليل نسبيا
الخوارزمية الجينية (GA)	لا تتضمن	صغير	كبير نسبيا
المنطق المضبب (FL)	تتضمن	كبير	قليل نسبيا

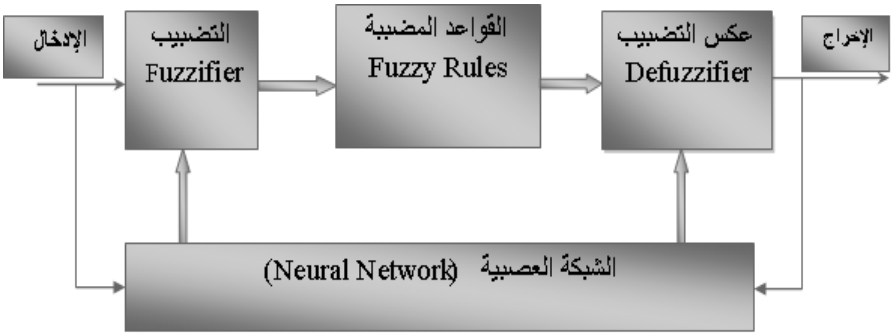
من الجدول (6) يتضح لنا أن كلاً من تقنيتي الشبكات العصبية والمنطق المضبب تحققان معيارين من مجموع المعايير الثلاثة, وهي عملية المحاكاة فضلاً عن الوقت القليل نسبياً في أثناء التطبيق, في حين تحقق الخوارزمية الجينية معياراً واحداً فقط وتفتقر إلى كل من محور المحاكاة الذي يعد الركن الأساس في التعرف على السلاسل الجديدة للحوامض الامينية عن طريق سلاسل الحامض النووي (DNA). فضلاً عن أنها تستغرق وقتاً أكبر من التقنيات الأخرى في عملية المعالجة, ولهذا استبعدناها وحاولنا التركيز على الشبكات العصبية والمنطق المضبب في عملية التهجين.

14- تهجين الشبكات العصبية الاصطناعية مع المنطق المضبب :

إن عملية تهجين الشبكات العصبية الاصطناعية مع المنطق المضبب تأخذ أشكالاً متعددة وطرقاً مختلفة, إلا أننا سوف نركز على عملية التهجين من خلال تعديل معلمات دالة العضوية بشكل يناسب بيانات الإدخال وذلك من خلال استخدام خوارزمية الانحدار المتدرج (Gradient Descent) التي يتم تطبيقها خلال عمل شبكة الانتشار الخلفي للخطأ (BP).

يتكون النموذج المهجن من خمس طبقات, ويستند إلى خوارزمية تعلم مهجنة لتمثيل معلمات أنظمة الاستدلال المضببة من نوع سوجينو (Sugeno), وهذه الخوارزمية تتضمن كلاً من طريقتي المربعات الصغرى (LS) و

شبكة الانتشار الخلفي (BP) وذلك من خلال مرحلتين من المعالجة. يتم في الأولى تطبيق طريقة المربعات الصغرى وتكون بالاتجاه الأمامي, ويتم في الثانية تطبيق خوارزمية شبكة (BP) وتكون بالاتجاه الخلفي, وهذا النموذج بأكمله يسمى معمارية نظام الاستدلال الضبابي العصبي المكيف (ANFIS).



الشكل (3) : يوضح معمارية نظام الاستدلال الضبابي العصبي المكيف (ANFIS).

إن تقنية (ANFIS) المهجنة كما في الشبكات العصبية الاصطناعية والمنطق المضبب تحتوي في معماريتها على عدد من الأجزاء التي يمكن اختيارها لإيجاد أفضل الحلول, ومن هذه الأجزاء الاختيارية تحديد عدد ونوع دوال العضوية المستخدمة في تمثيل بيانات الإدخال, ونفس هذا الاختيار موجود أيضا في أثناء بناء النموذج المضبب, كما تحتوي معمارية تقنية (ANFIS) على عملية المحاكاة التي من خلالها يتم إدخال بيانات جديدة.

15- اختيار عدد ونوع دوال العضوية المستخدمة في تقنية (ANFIS) :

إن الحصول على تقنية (ANFIS) المثلى لتصنيف بيانات سلاسل الحامض النووي (DNA) وتحويلها إلى الأحماض الامينية تتطلب اختيار أفضل المعلمات المستخدمة في بناء معمارية هذه التقنية. ومن هذه المعلمات, عدد ونوع دوال العضوية المستخدمة.

إن عملية مقارنة عدد دوال العضوية المستخدمة في تمثيل كل إدخال هي عملية صعبة وغير مقيدة بعدد معين, لذلك فقد قمنا بتثبيتها بأربعة دوال لكل إدخال ($p_i , i = 1,2,3$) من مصفوفة الإدخال ($P = [p1; p2; p3]$) وذلك بعد إجراء مقارنات متعددة تبين من خلالها أن أربعة دوال عضوية لكل إدخال تعطي أفضل تمثيل لعدد دوال العضوية في بناء معمارية التقنية المهجنة. أما أفضل اختيار لنوع دالة العضوية فهو دالة (Triangular) والتي تم مقارنتها مع بقية أنواع الدوال على النحو الآتي:

الجدول (7) : يوضح أفضل المعلمات المستخدمة في تمثيل تقنية (ANFIS).

الإدخال الأول (p1)	*الإدخال الثاني (p2)*	*الإدخال الثالث (p3)*	*معدل مربع الخطأ (MSE)*	*الزمن بالثانية (Time)*	*مقياس الكفاءة (EFF)*
Triangular	Trapezoidal	Gaussian	**6.293 e-13**	**2.179**	**7.292e+11**
Triangular	Gaussian	Trapezoidal	**6.378 e-13**	**2.388**	**6.565e+11**
Trapezoidal	Triangular	Gaussian	**6.218 e-13**	**1.905**	**8.442e+11**

8.071e+11	**1.956**	**6.334 e-13**	Triangular	Gaussian	Trapezoidal
8.666e+11	**1.818**	**6.347 e-13**	Trapezoidal	Triangular	Gaussian
9.137e+11	**1.751**	**6.250 e-13**	Triangular	Trapezoidal	Gaussian
9.305e+11	**1.807**	**5.947e-13**	Triangular	Triangular	Triangular
9.175e+11	**1.725**	**6.318 e-13**	Trapezoidal	Trapezoidal	Trapezoidal
6.034e+11	**2.091**	**7.925 e-13**	Gaussian	Gaussian	Gaussian

16- مقارنة بين التقنيات الذكائية المستخدمة في تصنيف الأحماض الامينية :

لقد تم مقارنة تقنية (ANFIS) المهجنة مع كل من المنطق المضبب (FL) والخوارزمية الجينية (GA) وشبكة الانتشار الخلفي للخطأ (BP) بأفضل حالاتها, وقد أثبتت النتائج أن التقنية المهجنة هي الأفضل من ناحية دقة النتائج المطلوبة التي يستدل عليها من خلال مقياس الكفاءة المتمثل بمعدل مربع الخطأ, وسرعة معالجة البيانات (الزمن المستغرق لمعالجة الإدخالات). والجدول (8) يوضح نتائج المقارنة:

الجدول (8) : مقارنة بين التقنيات الذكائية المستخدمة في تمثيل بيانات الحامض النووي (DNA).

نوع التقنية الذكائية المستخدمة *(Type of Technique)*	هل تتضمن عملية المحاكاة *(Simulation)*	معدل مربع الخطأ *(MSE)*	وقت التنفيذ بالثانية *(Time)*	مقياس كفاءة التقنيات *(EFF)*
التقنية المهجنة ANFIS	تتضمن	5.947e-013	1.807	9.305e+11
الشبكة العصبية BP	تتضمن	10.445	0.982	0.0975
المنطق المضبب FL	تتضمن	59.730	0.037	0.4525
الخوارزمية الجينية GA	لا تتضمن	7.5011e-5	8.661	1.539e+3

إن نتائج التهجين في الجدول (10) والمتمثلة بتقنية (ANFIS) أثبتت أنها هي الأفضل من خلال مقياس الكفاءة, فضلاً عن إمكانيتها على إجراء عملية المحاكاة, ثم تأتي بعد ذلك الخوارزمية الجينية في المرتبة الثانية من ناحية مقياس الكفاءة, إلا أن الخوارزمية الجينية لا تتضمن عملية المحاكاة, لذا لا يمكن استخدامها في تصنيف الأحماض الامينية, على الرغم من كفاءتها التي تفوق كل من الشبكة العصبية والمنطق المضبب.

17- الاستنتاجات والتوصيات :

إن التقنيات الذكائية المستخدمة في تصنيف الأحماض النووية أعطت نتائج متفاوتة, مما يثبت أن كل تقنية تتميز بمميزات رياضية خاصة بها, قد تصلح لتطبيقات معينة في حين تضعف أمام تطبيقات أخرى, مما يعطي شمولية للتطبيقات الحياتية المختلفة.

لقد تم مقارنة النتائج المستحصلة من الطرق الذكائية, وتبين من خلال المقارنة أن عملية التهجين بين الشبكات العصبية الاصطناعية والمنطق المضبب, تعطي نتائج أفضل عند تحويل سلاسل الحامض النووي (DNA) إلى الأحماض الامينية.

وبالنظر لما تتمتع به التقنيات الذكائية من مرونة في تركيبها, نوصي بإجراء عمليات تهجين بين هذه التقنيات للحصول على تقنيات محسّنة وتتمتع بمواصفات عالية.

المصادر:

[1]. Cartwright, H., (2008)**, " Using Artificial Intelligence in Chemistry and Biology: A Practical Guide"**, Taylor & Francis Group, LLC.

[2]. Jain, A.K., Duin, P.W. and Mao, J.,(2000) **, "Statistical Pattern Recognition: A Review"**, IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 22, NO. 1.

[3]. Keedwell, E. and Narayanan, A., (2005), **"Intelligent Bioinformatics: The application of artificial intelligence techniques to bioinformatics problems",** John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, England.

[4]. Steel, R.G.D. and Torrie, J.H., (1980), **"Principles and Procedures of Statistics a Biometrical Approach"**, Mc Graw-Hill, Inc.

[5]. Richards, J.E. and Hawley, R.S., (2005), **"The Human Genome: A users Guide"** , Second Edition, Elsevier Inc. All rights reserved.

[6]. Joshi, A. , Ramakrishman, N., Houstis, E.N. and Rice, J.R., (1997),