مجلة الدراسات التربوية والعلمية - كلية التربية - الجامعة العراقية
العدد العشرون - المجلد الخامس - علوم الفيزياء - كانون الأول 2022 م
doi.org/10.52866/esj.2021.05.20.10

نموذج ذكائي هجين مقترح لتحليل الانحرافات الجينية

ليلى هادي جوير ، شذى عبدالله محمد
جامعة الموصل / كلية علوم الحاسبات والرياضيات - قسم الرياضيات
Email: Layla.csp111@student.uomosul.edu.iq

مستخلص:

يمثل النمو السريع والحجم الهائل للبيانات الجينومية البشرية Genome Data ، خاصة تعدد الأشكال المنفرد للنمط الفردي Polymorphisms Nucleotide Single (SNPs)، تحديات للباحثين في الطب الحيوي فضلاً عن مبرمجي خوارزميات التعليم الآلي. أحدى هذه التحديات هي تحديد مجموعة فرعية مثلى من تعدد الأشكال SNPs، لايجاد معظم تنوع النمط المنفرد haplotype diversity لكل كتلة من النمط المنفرد haplotype block. تُسهّل عملية تقليل المعلومات هذه التنميط الجيني genotyping الفعال من حيث التكلفة، وبالتالي دراسة ارتباط النمط الجيني والنمط الظاهري. كما أن لها آثارًا على تقييم مخاطر تحديد موضوعات البحث على أساس معلومات SNP المتوفرة في بنك الجينات الالكتروني.

تم في هذا البحث اقتراح طريقة لاستخلاص الميزات المطلوبة لـSNPs من خلال اعتماد نواة جديدة في «خوارزمية تحليل المكون الأساسي "KPCA لغرض التمييز، وكذلك دمجها مع «خوارزمية البحث الذكائية تحسين سرب الجسيمات PSO»، وتم تسميتها PSO_KPCA. طبقت الطريقة المقترحة على خمس بيانات تجريبية حقيقة من موقع HapMap العالمي، ووجدت أن الطريقة المقترحة حققت دقة بنسبة 97.6٪ في حالة تمييز الـ SNPs المصابة. وتم تقدم الطريقة المقترحة في نموذج برمجي متكامل يتوافر على GUI ومكتوب بلغة Matlab®2016a، يقوم بقراءة ملف البيانات والي يكون بامتداد *.hmp، ويتم اجراء عدد من التحليلات الاستكشافية الشاملة للبيانات على مستوى الجينوم.

الكلمات المفتاحية: تعدد الاشكال المنفرد، النمط الجيني، النمط المنفرد، تحليل المكونات الاساسية، خوارزمية تحسين سرب الجسيمات.

A Proposed Hybrid Intelligent Model for Analyzing Genetic Deviations

Layla Hadi Jwear ، Shatha A. M Ramadhan
Email: Layla.csp111@student.uomosul.edu.iq

Abstract:

The rapid growth and massive volume of human genomic data, especially Single Nucleotide Polymorphisms (SNPs), presents challenges for biomedical researchers as well as programmers of machine learning algorithms. One such challenge is to identify an optimal subset of SNPs, to find the most haplotype diversity for each haplotype block. This information minimization process facilitates cost-effective genotyping, and thus the study of genotype-phenotype association. It also has implications for assessing the risk of identifying research subjects on the basis of SNP information available in the electronic gene bank.

In this paper, a method is proposed to select features for SNPs by adopting a new kernel in the KPCA "Primary Component Analysis Algorithm" for the purpose of discrimination, as well as combining it with the "Particle Swarm Optimization Intelligent Search Algorithm PSO", which is named PSO_KPCA. The proposed method was applied to five real experimental data from the global HapMap website, and it was found that the proposed method achieved an accuracy of 97.6% in case the affected SNPs were distinguished. The proposed method was presented in an integrated software model available on the GUI and written in Matlab®2016a, that reads the data file and has the extension *.hmp, and a number of comprehensive exploratory analyzes of the data are performed at the genome level.

Keywords: Single Nucleotide Polymorphisms, genotype, haplotype, PCA, PSO.

1 - مقدمة

الجينوم البشري هو المجموعة الكاملة لتسلسل الحمض النووي الريبي منقوص الأكسجين (DNA) Deoxyribonucleic Acid للبشر. يتكون من حوالي ثلاثة مليارات زوج قاعدي، مع أكثر من 99٪ من النيوكليوتيدات Nucleotides متطابقة تمامًا بين جميع السكان Population، مع فرق أقل من 1٪ بين الأشخاص. تحدث غالبية هذه الاختلافات الجينية على تعدد الأشكال المنفرد Single Nucleotide Poly-morphisms (SNPs). إذ تُعد SNPs أهم العلامات المستخدمة لرسم خرائط الأمراض بالجينات. على الرغم من أن معظمها محايد، فقد أظهرت الدراسات الحديثة أن بعض أشكال النيوكليوتيدات SNPs لها وظيفية تؤثر على النمط الظاهري، مثل الطول، ولون الجلد، والمقاومة، والعدوى، أو الاستجابة للأدوية، وما إلى ذلك[1].

الميزة الرئيسية التي تجعل SNPs مفضلة على التعبيرات الجينية الدقيقة هي الاستقرار عالي التردد والذي يكون أسهل وأسرع في التجميع [2]. في هذا السياق، تم تطبيق العديد من خوارزميات التعلم الآلي على نطاق واسع لتصنيف بيانات SNPs. ومع ذلك، فإن «عدد الأبعاد الكبير» هو التحدي الرئيسي الذي واجهته معظم الدراسات، نظرًا لأن عدد العينات samples (على سبيل المثال بضع مئات) أصغر بكثير من عدد الـ SNPs والتي تصل إلى قرابة مليون نيوكلويتد [1].

في الآونة الأخيرة، أصبح من المعترف به على نطاق واسع أن تفاعلات SNPs عالية المستوى قد تساهم في إحداث إمراضية معينة. إذ تمثل تفاعلات SNPs عالية الترتيب k-order (k>3) مزيجًا من تعدد الأشكال المنفرد SNPs التي تؤثر بشكل مشترك على الأمراض المعقدة سواء بشكل خطي أو غير خطي، ويكون البحث عنها وايجادها أهمية كبيرة لاكتشاف الأسباب المسببة للأمراض البشرية المعقدة.

تتمثل النقطة المركزية في الاكتشاف في كيفية التمييز بدقة في العلاقات بين إجراءات تداخل SNPs عالية الترتيب وحالات المرض وكيفية استكشاف تفاعلات SNPs عالية الترتيب بسرعة على نطاق واسع في الجينوم، حيث يكون التحدي الأكبر هو تطوير طريقة بحث فعّالة لحل الانفجار الاندماجي لـ SNPs. يتطلب البحث الشامل الذي تم تطبيقه على نطاق واسع للعثور على تفاعلات SNPs الزوجية تكلفة افتراضية كبيرة لتقييم الارتباط بين جميع مجموعات SNPs، وهو أمر غير عملي لاكتشاف تفاعلات SNPs عالية الترتيب على على نطاق الجينوم.

لمعالجة هذه المشكلة، تم اقتراح بعض تقنيات البحث، مثل الحوسبة عالية الأداء وعمليات البحث العشوائية، لتسريع اكتشاف تفاعلات SNPs عالية الترتيب. تعتمد الحوسبة عالية الأداء بشكل عام على أجهزة الحاسوب العملاقة أو تقنيات المعالجة المتوازية، مثل الحوسبة العنقودية Cluster Computing System، والحوسبة المتوازية Parallel Computing System، لتسريع العمليات الحسابية.

تم تطوير العديد من الخوارزميات للبحث الأمراض المعقدة وتصنيفها بناءً على مجموعات بيانات SNP. الباحث Evans وآخرون [3] استخدموا طريقتين هما طريقة الفرز الفردي والخوارزميات الإحصائية المعيارية لمربع كاي Chi-squared، جنبًا إلى جنب مع حاسوب المتجهات الداعمة Support Vector Ma- (SVM) chines مع وظائف نواة متعددة كطريقة تصنيف، تصل إلى دقة تصل إلى 73٪.

أما الباحث Batnyam وآخرون [4] فقد جمعوا العديد من تقنيات الخدمة الثابتة الحالية: نهج من ثلاث مراحل **أولاً**، اختيار SNPs الأكثر إفادة باستخدام اختيار الميزة على أساس قيمة R (RFS) [5]، اختيار الميزة على أساس تمييز المسافة (FSDD) [6]، الميزة الإغاثة القائمة على الوزن [7] وخوارزمية تعتمد على وضوح الميزة (CBFS) [8]. **ثانيًا**، إنشاء ميزة اصطناعية من SNPs المحددة باستخدام طريقة دمج الميزات (FFM)، و**ثالثًا**، التصنيف باستخدام مصنف مصنع الجينات الاصطناعية (AGM) Artificial Gene Making وSVM ومصنف الجيران الاقرب k-Near-est Neighbor (k-NN). تم تحقيق أفضل دقة لجميع مجموعات البيانات المختبرة باستخدام مزيج من CBFS وFFM، المصنفة باستخدام SVM. تم تطبيق الطريقتين المذكورتين أعلاه على مجموعتي بيانات التخلف العقلي (MR) واضطراب طيف التوحد (ASD).

كما حلل Goudey وآخرون [9] تفاعلات SNPs عالية الترتيب k-order (k>3) باستخدام الطريقة الشاملة Exhaustive method باستخدام حاسبات ذات الاداء العالي. أما المؤلفون Wan وآخرون، وYung وآخرون و Gyenesei وآخرون، فقد وظفوا عمليات منطقية على مستوى أحاديات المعامل مبرمجة بشكل متوازي لتحقيق كفاءة حسابية ممتازة [10]–[12]. إذ تم استخدام البحث العشوائي لإجراءات أخذ العينات العشوائية للعثور على مجموعات SNPs عالية الترتيب المرتبطة بحالة المرض. وتهدف هذه الطريقة إلى تقليل التعقيد الزمني عن طريق تقليل عدد تركيبات SNPs اللازمة لحساب الارتباط مع النمط الظاهري.

كما اقترح الباحثان Zhang و Liu طريقة مونت كارلو لسلسلة ماركوف (MCMC) للكشف المتكرر عن التفاعلات المعرفية عن طريق حساب احتمالية الارتباط الخلفي للموضع في التفاعل مع المرض [13]. أما الباحثون Wang وآخرون، فقد استخدموا طريقة MCMC للكشف عن تفاعلات SNP عالية الترتيب [14]. وطبق المؤلفون Han وآخرون، خوارزمية سريعة ذات تفرع ومنضم و MCMC لفحص تعدد الأشكال [15]. أما الباحثون في [16] فقد استخدموا خوارزمية SNPHarvester في اجراء بحثًا عشوائيًا لاكتشاف مجموعات SNPs المهمة.

تحاول الحوسبة عالية الأداء تقييم ارتباط جميع مجموعات SNPs ذات الترتيب k باستخدام أنظمة الكمبيوتر عالية الأداء، مثل الحوسبة المتوازية والحوسبة السحابية. ومع ذلك، عندما يكون الترتيب كبيراً يعني k>3، لا يزال اكتشاف تفاعلات SNP بترتيب k من البيانات الجينية مع مئات الآلاف من SNPs غير فعال بسبب العبء الحسابي الهائل.

فعلى الرغم من أن خوارزميات البحث العشوائية التقليدية يمكنها اكتشاف بعض تفاعلات SNPs ذات الترتيب k، إلا أن هذه الخوارزميات لا تزال غير كافية للكشف عن تفاعلات SNPs عالية الترتيب عند استخدام احتمالات الارتباط اللاحق للمواقع الفردية، خاصةً للكشف عن نماذج الأمراض المعقدة ذات التأثيرات الهامشية الدنيا أو المعدومة.

في هذا البحث تم تقديم نموذج ذكائي هجين وهو عبارة عن حزمة برامج مكتوبة في Matlab لتحليل البيانات في علم الوراثة السكانية الجزيئية. باعتبارها لغة عالية الأداء للحوسبة التقنية، فقد حظيت Mat-lab بتقدير متزايد من قبل علماء الأحياء لتحليل البيانات [17]. ومع ذلك، يتوافر عدد قليل من دوال Matlab لتحليل بيانات التباين في التسلسل الجيني (بما في ذلك تعدد الأشكال المنفرد تسلسل الحمض النووي

وتعدد الأشكال المنفرد [SNPs]). تم تطوير البرنامج المقترح لملء هذه الفجوة، وتوفير دوال لمعالجة البيانات، وحساب الإحصاءات الجينية للسكان.

يتميز البرنامج المقترح بالسهولة في الاستخدام، وذلك لتوفره على الواجهات والمخرجات المطورة. وهو قابل للتطوير بدرجة كبيرة لأنه يمكن إعداده بسهولة (استدعاء دالة تلو الأخرى) لأداء مهمة كاملة في وضع الدُفعات غير المراقب. علاوة على ذلك، فهو يحتوي على واجهات مستخدم رسومية بسيطة وفعالة قائمة على القوائم وموجهة إلى مربعات الحوار، والتي تخفي تعقيد العمليات الحسابية عن المستخدمين. فضلاً عن عمله تحت جميع أنظمة التشغيل الثلاثة الرئيسية MSWindows وLNIX وMac. يتميز الإصدار الأحدث من Matlab بكفاءة الاستخدام الامثل للذاكرة، مثل الحساب الأحادي الدقة، ووظيفة تعيين الذاكرة، ودعم الأنظمة الأساسية 64 بت. فعند استخدام هذه الميزات، فان البرنامج المقترح يكون قادرًا على التعامل مع مجموعات بيانات علم الوراثة الحديثة على نطاق آلاف الأفراد ومئات الآلاف من المتتابعات أو SNPs.

تم تنظيم بقية هذه الورقة في أربعة أقسام عدا القسم الاول: القسم التالي تم فيه استعراض بعض المصطلحات البيايولوجية المستخدمة في هذه الدراسة، بعده يأتي وصف الإطار النظري والعملي للبرنامج المقترح في القسم الثالث. ومن ثم يناقش القسم الرابع التقييم التجريبي والنتائج التي تم الحصول عليها، بينما يتم استخلاص النتائج في القسم الخامس.

2 - الانحراف الجيني Genetic Drift:

يمثل التغيير الجيني الحاصل في مجتمع سكاني، أو كتلة سكانية صغيرة نسبياً، بحيث يأخد شكلاً حازماً في التغيير الواقع مصادفة، فيعمل هذا التغيير الجيني على تغير في صفات جينية معينة؛ من حيث انتشارها بشكل واسع أو قليل، حسب موقع حدوثها، وأهمية ذلك التغيير الجيني، وشكل تعبيره جينياً، وظهوره على الصفات الشكلية[18].

لقد تطور اليوم إلى نظام يعتمد على البيانات، إذ تختبر مجموعات البيانات الجينومية واسعة النطاق حدود النماذج النظرية وطرق التحليل الحسابي. تحليلات بيانات تعدد الأشكال المنفرد SNPs في تسلسل الجينوم الكامل من البشر والعديد من الكائنات الحية النموذجية تعطي رؤى جديدة فيما يتعلق بتاريخ السكان والانتشار الجيني للانتقاء الطبيعي.

سيتم في هذا المبحث عرض لأهم التعاريف البيايولوجية المستخدمة في هذه الدراسة، لفهم الانحرافات الجينية بشكل متكامل.

تعريف (1): الموروثات أو الجينات[19]:

هي الوحدات الأساسية للوراثة في الكائنات الحية وضمن هذه الموروثات يتم تشفير المعلومات المهمة لتكوين أعضاء الجنين والوظائف العضوية الحيوية له، وتتواجد عادة ضمن المادة الوراثية التي تمثلها الـ DNA.

تعريف (2): الأليل أو الحليل أو البديل[20]:

الأليل Allele أو الحليل هو نسخة أو شكل بديل للجين أو الموقع الجيني (عادة يتكون من مجموعة جينات)، وله نسختان أو شكلان بديلان على الأقل. أحيانا قد ينتج عن ألألائل المختلفة المحتوية على اختلافات في الشفرة الوراثية صفات مختلفة (كلون الجلد أو العين). إلا أن الكثير من الإختلافات لا تؤدي إلا لاختلاف ظاهري طفيف أو معدوم[21].

تنتقل المادة الوراثية من جيل لآخر، خلال عملية

التكاثر Reproduction، بحيث يكتسب كل فرد جديد نصف مورثاته من كروموسومات أحد والديه والنصف الآخر من كروموسومات الوالد الآخر، هذا يعني أنها تحمل نسختين من كل جين (أي زوجاً من الألائل لكل جين). إن كان كلا الأليلان متطابقان، فالكائن الحامل لها يسمى زايگوت متماثل Homozy-gous، ويصف بمتماثل الزايگوت. وإن كان الأليلان مختلفان، فالحامل لها هو زايگوت متباين ويصف بمتباين الزيجوت Heterozygous.

يمتلك الأفراد المنتمون لتجمع سكاني Popula-tion من الكائنات الحية عدة ألائل (نسخ مختلفة من الجينات) لكل موضع كروموسومي. ويمكن قياس درجة التنوع الأليلي لموضع كروموسومي معين بعدد الألائل الموجودة (تعدد الأشكال)، أو بالنسبة للاشكال التي يشكلها متباينو الزيجوت من التجمع[21].

تعريف (3): علم الوراثة السكانية Population

يفحص علم الوراثة السكانية تكرارات وتوزيع الأليات والأنماط الوراثية والنمط المنفرد في المجموعات الطبيعية والاصطناعية والمحاكاة. الآليات (الانتقاء الطبيعي، الطفرة، الانحراف الوراثي، التزاوج التجميعي، الهجرة) لتشكيل التباين الوراثي داخل المجموعات وبينها، وعلاقتها بالتطور، وعلم النظم، وعلم التطور، والأساس الجيني للصفات، مثل سبب المرض الوراثي، والقائم على الحمض النووي تعتمد طرق تحديد الصفات الوراثية على فهم أنماط التنوع في المجموعات البشرية.

تعريف (4): تعدد الأشكال المنفرد

Polymorphisms Nucleotide Single SNPs

تعدد الأشكال المنفرد هو اختلاف في متتابعة الـ DNA، يحدث عادة في اي تجمع سكاني، إذ يختلف نيوكليوتيد واحد A، C، G أو T في الجين بين فردين من نفس النوع البايولوجي أو بين كروموسومات مزدوجة[22].

مثال: ادناه مقطعي متتابعة لـ DNA من شخصين مختلفين:

A A G C C T A

A A G C T T A

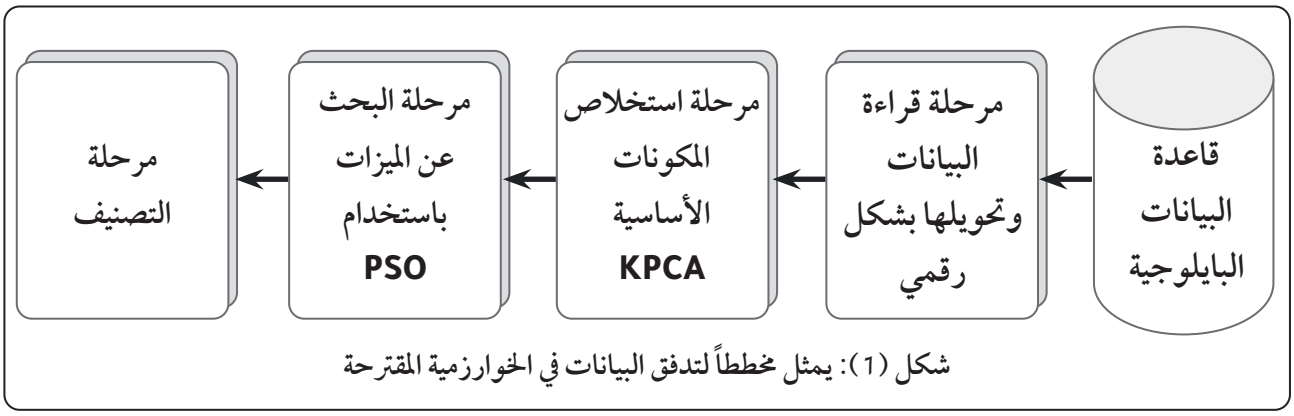
في هذا المثال الـSNP يمثل اختلافاً في نيوكليوتيد واحد فقط.

قد يقع تعدد الأشكال المنفرد النوكليوتيد المفرد ضمن متتابعات الترميز في الجينات، أو مناطق غير الترميز، أو في المناطق بين الجينات. تعدد ألأشكال داخل تسلسل الترميز لا يغيّر بالضرورة تسلسل الأحماض الأمينية في البروتين الناتج، وذلك بسبب صفة الانحلالية في الشفرة الوراثية[23].

3 - الخوارزمية الهجينة المقترحة:

في هذا المبحث سيتم اقتراح خوارزمية هجينة لتحليل الانحرافات الجينية لبيانات تحتوي أنماطاً جينية Genotypes. إذ تعتمد الطريقة المقترحة على الاستفادة من خوارزمية تحليل المكون الاساسي PCA، المستندة على النواة K، لتقليل ابعاد البيانات المطلوب تحليلها، ومن ثم تمييزها باستخدام خوارزمية تحسين سرب الجسيمات PSO.

مراحل الخوارزمية المقترحة لتحليل الأمراض المعقدة موضحة في الشكل (1) وكما يلي: **أ)** مرحلة ما قبل المعالجة تتكون من قراءة البيانات وتحويلها بشكل رقمي، **ب)** مرحلة استخلاص المكونات الاساسية باستخدام خوارزمية KPCA المستندة على النواة، **ج)** مرحلة البحث عن الميزات باستخدام الخوارزمية الذكائية PSO، و **د)** مرحلة التصنيف.



شكل (1): يمثل مخططاً لتدفق البيانات في الخوارزمية المقترحة

فيما يلي توضيح للخطوات اعلاه:

3 – 1. مرحلة قراءة البيانات وتحويلها

هذه المرحلة تنقسم إلى مرحلتين فرعيتين: قراءة البيانات بالتنسيق المحدد بوساطة مشاريع HapMap [24]، والتي تكون عادةً كمتتابعات نيوكلويتدات وبالتالي فهي ليست دائمًا مناسبة للتحليل، لذلك، يجب تحويلها إلى شكل رقمي، وهذه هي المرحلة الثانية. عملية التحويل تكون بشكل مباشر، ويمكن تطبيقها بطرق مختلفة.

3–2. مرحلة استخلاص المكونات الاساسية KPCA

تعد معالجة البيانات وتقليل ابعادها خطوة أساسية قبل إجراء أي تحليل رياضي للبيانات، مع الحفاظ على الكثير من المعلومات المهمة التي لها صلة بالبيانات الأصلية والتي تعد من المتطلبات الأساسية في التصنيف[25]. فمعظم البيانات تحتوي على مجموعة جزئية من الميزات غير الضرورية أو بيانات مكررة، وبالتالي تؤثر سلباً على عملية حساب الميزات من خلال استغراق وقت أطول في التنفيذ بالإضافة الى الكلفة العالية [25].

لهذا توجب استخدام طرق وأساليب رياضية مثل نواة تحليل المكون الأساسي Kernel Principal Component Analysis (KPCA) الذي يُعدّ أحد أفضل أساليب البحث عن ميزة معينة او صفات مميزة من خلال تحويل ابعاد البيانات الأصلية من مساحة عالية الابعاد إلى مساحة مميزة ذات بعدٍ أقل، من خلال اختيار الأفضل من بين مجموعة من الميزات الأصلية الى مجموعة فرعية تضم ميزات اساسية وجوهرية مطلوبة في عملية التمييز[26].

3–2–1 نواة تحليل المكون الأساسي Kernel Principal Component Analysis (KPCA)

تستند طريقة نواة تحليل المكون الأساسي (KPCA) في تحويل البيانات من الفضاء الأولي إلى فضاء جديد ذي بُعد أكبر، وذلك من اجل اعتماد التطبيق غير الخطي على البيانات. حيث تكون العلاقة بين المتغيرات في هذا الفضاء الجديد أكثر وضوحًا على عكس العلاقة الأولى وبالإضافة الى زيادة عدد الابعاد. ان تقنية وأساسيات استخدام أساليب النواة، تمكننا من الاستفادة من خدعة النواة والغرض هو لاشتقاق نسخة غير خطية من تحويل كارنولوف Karhunen Loeve Transforma-tion (KLT) الذي يتعامل مع البيانات الخطية فقط وهذه مشكلة في كثير من البيانات. واصبحت فكرة الاستفادة من النواه من خلال عمل نسخة غير خطية او بمعنى تكوين امتداد غير خطي للبيانات الخطية لكي يتم التعامل مع البيانات بكل حرية[27].

ان الهدف الرئيس في استخدام خوارزمية (KPCA) هو توفير حلاً واحًدا في كيفية العثور على المساحة البعدية المخفضة المناسبة حال كون البيانات غير خطية. ويتم ذلك من خلال تشكيل البيانات إلى مساحة

أعلى ذات أبعاد غير خطية. وبالمقارنة مع الطريقة الكلاسيكية وجد انها توفر معدل تصنيف أفضل من تحويل (KLT) الخطية، إذ يمكن استخلاص المزيد من البيانات والمميزات باستخدام هذه الطريقة مقارنة مع تحويل(KLT) الخطية التقليدية.

3-2-2. خطوات خوارزمية نواة تحليل المكون الاساسي (KPCA)

تم تطبيق خوارزمية نواة تحليل المكون الأساسي (KPCA) بالشكل الآتي:

خوارزمية نواة تحليل المكون الأساسي KPCA

المدخلات: مجموعة بيانات SNPs، والتي هي عبارة عن ملف *.hmp يحتوي على N من الافراد، ولكل فرد M من الـSNPs.

المخرجات: مصفوفة المسافة لبيانات التدريب والاختبار .

1: اجراء المعالجة الأولية على البيانات المدخلة، ويتم تقسيمها الى بيانات تدريب وبيانات اختبار.

2: حساب مصفوفة التباين لبيانات التدريب من خلال من المعادلة:

$$\lambda_k V_k = C V_k$$

حيث $\lambda_k \geq 0$ هي القيم الذاتية، والمتجهات الذاتية غير الصفرية. $V_k \in \frac{\psi}{\{0\}}, V_k^T V_k = 1$

3: إيجاد كل من القيم الذاتية والمتجهات الذاتية من خلال المعادلتين:

$$K_{ij} = K(X_i, X_j) \ , \ M\lambda_k K a_k = K a_k$$

4: استخلاص الميزات من خلال استخدام معادلة النواة :

$$K_{pol}(x_i, x_j) = \left(c + x_i^t x_j\right)^p, \quad p \geq 1$$

حيث c نواة التباين الناتجة من مصفوفة النواة.

5: حساب مركز النواة من خلال المعادلة:

$$\tilde{K} = K - QK - KQ + QKQ$$

حيث Q هي مصفوفة $(M \times M)$ جميع الادخالات التي تساوي $(\frac{1}{M})$.

6: ترتيب الحلول واستخلاص الميزات المطلوبة من المتجهات الذاتية الموضحة في الخطوة (7) من خلال المعادلة السابقة.

7: تكوين نموذج (Model) يحوي على الميزات المستخلصة، والانتهاء من لبيانات التدريب.

8: اسقاط بيانات التدريب المستخلصة من خوارزمية (KPCA) على بيانات الاختبار $Z(x_2)$ لغرض تهيئة بيانات الاختبار للتمييز.

9: اجراء اختبار بين بيانات الناتجة من خوارزمية (KPCA) أي بين بيانات التدريب وبيانات الاختبار باستخدام قانون المسافة الاقليدية:

$$dRT = DRT(x_R, x_T)$$

حيث x_T ،x_R تمثلان بيانات التدريب والاختبار، على التوالي، الناتجتين من خوارزمية (KPCA).

3-3. مرحلة البحث عن الميزات باستخدام PSO

تعد خوارزميات تحسين سرب الجسيمات (PSO) Particle swarm optimization احدى التقنيات المستخدمة في نظام المعلومات الجغرافية GIS، طورها العالمان Eberhart and Kennedy. تحاكي هذه الطريقة سلوك البحث عن الغذاء لسرب من الطيور أو الأسماك [28]. إذ يُشار إلى كل جسيم على أنه طائر أو سمكة وله سلوكان: اجتماعي وفردي. تسترشد حركات الجسيمات بأفضل مواقفها المعروفة مسبقاً (السلوك الفردي)، والموقع الأكثر معرفة للسرب الحالي (السلوك الاجتماعي)، والذي يمكن التعبير عنه على النحو التالي:

$$V_i(t+1 = \omega V_i(t) + rand(0,1) \times c_1 \times \left(X_i^{old} - X_i^{best}\right) + rand(0,1) \times c_2 (X_i^{old} - X^{Gbest}(t))$$

$$X_i(t+1) = X_i(t) + V_i(t+1)$$

حيث تمثل:

$(X_i^{old} - X^{Gbest}(t))$ و $(X_i^{old} - X_i^{best})$

السلوك الاجتماعي والسلوك الفردي، على التوالي، للجسيم X_i ، ويشير c_1 و c_2 إلى عوامل التعلم للجسيم X_i. اما فتمثل $\omega \in (0,1)$ وزن القصور الذاتي لسرعة التعلم. X_i^{best} هو أفضل موقع سابق للجسيم i . ويشير X^{Gbest} إلى أفضل موقع شاملاً للسرب بأكمله.

تتميز خوارزمية PSO بمعدل تقارب مرتفع جدًا لحل بعض مشكلات الأمثلية المعقدة. Yang وآخرون، قدموا خوارزمية PSO لأول مرة والتي سميت الخريطة المزدوجة الفوضوية (DBM-PSO) Double-bottom chaotic map PSO للكشف عن التفاعلات الجينية عالية الترتيب [48]. إذ تم في هذه الخوارزمية اعتماد خرائط مزدوجة القاع لموازنة قوة الاستكشاف، بهدف منع PSO من أن يصبح محاصرًا في المستوى المحلي الأمثل. يشير الجسيم إلى حل مرشح يمثل تفاعل SNPs بترتيب k. أما Shang وآخرون، فقد اقترحوا خوارزمية لتحسين التعلم القائم على تباين PSO (والمسمى IOBLPSO) للكشف عن تفاعلات SNP-SNP [1]. إذ يتم استخدام IOBLPSO التعلم القائم على التباين لتحسين قوة الاستكشاف الشاملة، ويستخدم وزن القصور الذاتي الديناميكي لتغطية البحث الواسع، إذ تعتمد المعالجة اللاحقة لإجراء بحث عميق في مجموعات SNP المشتبه بها.

التعقيد الزمني لطريقة PSO للكشف عن تفاعل SNP بترتيب k يكون بحساب مجموع كل من الخطوات المذكورة في الجدول (1).

الجدول (1): مجموع التعقيد الزمني لطريقة PSO

ت	الخطوة	التعقيد
1	القيم الابتدائية	$O(k \times NP)$
2	تحديث السرعة	$O(T \times k \times NP)$
3	تحديث الموقع	$O(T \times k \times NP)$

تعد PSO طريقة فعالة للغاية لحل مشاكل الأمثلية المستمرة وإيجاد منطقة جينية صغيرة مستمرة تحتوي على تعدد الأشكال المتعددة المسببة للأمراض. وهنا لا بد من التأكيد على ثلاثة تفاصيل مهمة، لم يتم ذكرها في الاعمال السابقة في تصميم خوارزمية الكشف باستخدام PSO:

1) للحصول على تركيبة SNPs ،

$$X = (x_1, x_2, \dots, x_n)$$

يجب أن تكون متغيرات القرار x_i عددًا صحيحًا. ومع ذلك، يمكن أن يكون مشفرًا بقيمة حقيقية لتسريع البحث.

2) إذا كان $x_i \neq x_j$. $i < j$

3) اذا كان $x_i < x_j$. $i < j$. وهذا الشرط مهم لتحسين القدرة على البحث وتجنب العقبات أثناء البحث المحلي.

تم تصميم جميع خوارزميات PSO بناءً على هذه القواعد، والتي يمكن أن تحسن أداء الخوارزميات بشكل فعال.

3-4. مرحلة التصنيف

تهدف خوارزميات التعلم الآلي إلى تطوير برامج كمبيوتر قادرة على التعلم من تلقاء نفسها واكتشاف الأنماط في البيانات، وتغيير إجراءات البرنامج وفقًا للبيانات الجديدة. تم استخدام تقنيات مختلفة للتعلم الآلي بخصائص مختلفة في هذه الدراسة. الغرض

الرئيس من استخدام هذه المصنفات هو قياس أداء طريقة اختيار الميزات المختلطة المقترحة.

4 - النتائج

في هذا المبحث سيتم استعراض الخطوات التي تعمل بها الخوارزمية المهجنة المقترحة ومناقشتها وتطبيقها. تم كتابة الكود البرمجي باستخدام (Matlab®2016a) لما تتميز بها هذه البرمجة كلغة حسابية علمية قوية ومريحة في علم الوراثة السكانية الجزيئي. إذ يمتلك مجموعة أدوات قوية ومرنة مخصصة لتحليل تعدد الأشكال المنفرد تعدد الأشكال SNPs. فضلاً عن تطبيق عددًا من الخوارزميات والأساليب والأدوات لإجراء تحليلات استكشافية شاملة للبيانات على مستوى الجينوم.

4-1. مجموعات بيانات SNPs

ان الطريقة العامة لقياس مدى كفاءة أي نظام مقترح باستخدام التقنيات الذكائية تعتمد على مرحلتين

1. مرحلة التدريب: وهي عملية بناء نماذج بحسب نوع الميزات المطلوبة في عملية التمييز.

2. مرحلة الاختبار: وهي عملية اختبار النظام من خلال بيانات مخصصة للاختبار مع القوالب والنماذج المخزونة في مرحلة التدريب لمعرفة مدى كفاءة النظام.

في هذا العمل، تم الاستعانة بالموقع العالمي NCBI والذي يمثل مستودعاً عاماً الكترونياً، يقوم بأرشفة وتوزيع تسلسل الجيل التالي Next-Generation والبيانات الجينومية الأخرى بحرية، فضلاً عن المشروع HapMap الذي يحوي بيانات SNPs لجميع مجموعات HapMap الإحدى عشر: ASW: أصل أفريقي في جنوب غرب الولايات المتحدة الأمريكية، CEU: سكان يوتا من أصول أوروبية شمالية وغربية من مجموعة CEPH، CHB: هان الصينية في بكين/ الصين، CHD: صيني في متروبوليتان دنفر/ كولورادو، GIH: هنود الغوجاراتية في هيوستن/ تكساس، JPT: ياباني في طوكيو/ اليابان، LWK: Luhya في Webuye/ كينيا، MXL: أصل مكسيكي في لوس أنجلوس/ كاليفورنيا، MKK: الماساي في كينياوا/ كينيا، TSI: توسكاني/ إيطاليا، YRI: اليوروبا في إبادان/ نيجيريا[29]. تتكون مجموعات البيانات من عينات مصنفة كحالة للأفراد المتضررين أو للتحكم في الأصحاء، ولكل منها رقم تعريف وتسلسل من أليلات SNPs. يمكن أن تكون SNPs متغايرة الزيجوت أو متماثلة اللواقح.

يطرح تطبيق تقنيات التعلم الآلي على مجموعات البيانات هذه بعض التحديات، إذ يُعدّ حجم العينة الصغير أحد أكثر المشكلات صعوبة في تحليل مجموعة بيانات SNPs. فيؤثر العدد الصغير للعينات والعدد الهائل من SNPs بشكل كبير على تقدير الخطأ. وتجدر الإشارة إلى أن اختيار طريقة التحقق الصحيحة أمر ضروري لتقدير خطأ التصنيف. العدد القليل من SNPs يؤدي إلى زيادة احتمالية التكرار في السمات، ويجعل الخوارزمية المقترحة معقدة نوعاً ما[22].

4-2. اختبار النتائج

أجريت تجارب تصنيف خاضعة للإشراف لتقييم أداء الخوارزمية المقترحة. تم تطبيق حساب المسافة الاقليدية Euclidean Distance (ED) على بيانات التدريب والاختبار، على التوالي. فتم تقييم خوارزمية KPCA ولكن أدائها كان دون المستوى مقارنة بالخورازمية المقترحة التي تم اختبارها. تم تنفيذ الخوارزميات بالكامل باستخدام (Matlab®2016a). وتم تقييم معدل التنبؤ باستخدام مقياسين، متوسط الدقة Average Accuracy (Acc) وكما موضح في أدناه:

$$Acc = \frac{TP+TN}{TP+FN+FP+TN} \quad \ldots\ldots\ldots\ldots (1)$$

ومقياس F-measure (F) المعادلة (2)، حيث تمثل قيمة F قدرة الطريقة المقترحة على التنبؤ بحالات الأفراد المتضررين.

$$F = 2.\frac{Pre.Re}{Pre+Re} \quad \ldots\ldots\ldots\ldots\ldots (2)$$

حيث ان:

$$Re = \frac{TP}{TP + FN} \times 100$$

$$Pre = \frac{TP}{TP + FP} \times 100$$

يشير كل من TP و FN و FP و TN إلى عدد المتنبئين الموجب الحقيقي، والسلبي الخاطئ، والإيجابي الخاطئ، والسلبي الحقيقي، على التوالي.

4-3. الخوارزميات المستخدمة

تم في الفصول السابقة التعرف على الجانب النظري للخوارزميات المستخدمة، تتضمن الفقرات الآتية عرض نتائج الخوارزميات المستخدمة واجراء مقارنة بين النتائج والتي سيرد ذكرها لاحقاً.

4-3-1. خوارزمية نواة تحليل المكون الأساسي KPCA

ان السبب الأساسي للاهتمام باستخدام خوارزمية (KPCA) هو في استخلاص الميزات غير الخطية حيث بعد اجراء المعالجة الأولية تم تقسيم البيانات الى بيانات تدريب وبيانات اختبار. وتم اجراء عملية استخلاص الميزات بيانات التدريب من خلال خوارزمية (KPCA) واجراء تهيئة بيانات الاختبار لغرض اجراء اختبار بين بيانات التدريب والاختبار. وتم اختبار النتائج والحصول على افضل النتائج عند استخدام المجموعة CEU للكروموسوم رقم 13، والتي كانت نسبة الرفض تمثل (6.5٪)، اما نسبة القبول فكانت (93.5٪)، وكما موضح في الجدول (2).

الجدول (2): يمثل نسب الرفض والقبول لخوارزمية KPCA

البيانات المستخدمة		ACC	F*	نسبة القبول	نسبة الرفض
المجموعة	رقم الكروموسوم				
ASW	Chr10 :68983..135327873	69.44	69.76	91.2%	8.8%
CEPH	Chr4 :119204834..164186769	69.76	69.44	91.2%	8.8%
CEPH	Chr21 :16713718..44997989	75.31	76.45	92.5%	7.5%
CEPH	Chr11 :34647271..91409074	63.16	65.28	90.9%	9.1%
CEU	Chr13 :31787616..31871804	85.00	85.21	**93.5%**	**6.5%**
CEU	Chr12 :77066248..77066248	63.17	65.68	91.8%	8.2%
CHB	Chr10 :88481..135284978	73.92	76.67	91.2%	8.8%
JPT	Chr2 :9715614..9721513	76.11	76.92	92.1%	7.9%
YRI	Chr2 :9715614..9721513	60.46	65.28	90.3%	9.7%
YRI	Chr23 :23700229..50360164	70.52	74.08	91.2%	8.8%

* ملاحظة: قيمة F تمثل قدرة الطريقة المقترحة على التنبؤ بالحالات.

4-3-2. الخوارزمية المهجنة المقترحة:

تم تهجين خوارزمية تهدف الى دقة عالية في التمييز، إذ تم ادخال ملفات hmp.* مع اجراء المعالجة الأولية على البيانات وتقسيمها الى بيانات تدريب وبيانات اختبار، ومن ثمّ اجراء عملية استخلاص الميزات باستخدام خورازمية نواة تحليل المكون الأساسي (KPCA) مع اعتماد نواة ($K_{\text{polonomial}}$) وتهيئة بيانات الاختبار ومن ثم اجراء اختبار بين بيانات التدريب مع بيانات الاختبار، بعدها يتم اجراء اختيار للميزات باستخدام خوارزمية بحث PSO على بيانات التدريب. الجدول (5) يوضح ارتفاع نسبة التمييز بدرجة ملحوظة، إذ تم الحصول على (4.3٪) التي تمثل نسبة الرفض، والحصول أيضا على نسبة (95.7٪) التي تمثل نسبة القبول عند استخدام الملفات ذاتها.

الجدول (3): يمثل نسب الرفض والقبول للخورازمية المهجنة المقترحة

البيانات المستخدمة		ACC	F*	نسبة القبول	نسبة الرفض
المجموعة	رقم الكروموسوم				
ASW	Chr10 :68983..135327873	72.91	73.25	92.5%	9.5%
CEPH	Chr4 :119204834..164186769	73.25	72.91	92.7%	7.3%
CEPH	Chr21 :16713718..44997989	79.08	80.27	93.9%	6.1%
CEPH	Chr11 :34647271..91409074	66.32	68.54	91.2%	8.8%
CEU	Chr13 :31787616..31871804	87.91	87.59	**95.7%**	**4.3%**
CEU	Chr12 :77066248..77066248	82.75	83.14	93.5%	6.5%
CHB	Chr10 :88481..135284978	77.62	80.50	93.1%	6.9%
JPT	Chr2 :9715614..9721513	84.00	86.03	94.7%	5.3%
YRI	Chr2 :9715614..9721513	68.54	66.32	90.1%	9.9%
YRI	Chr23 :23700229..50360164	74.05	77.78	92.4%	7.6%

* ملاحظة: قيمة F تمثل قدرة الطريقة المقترحة على التنبؤ بالحالات.

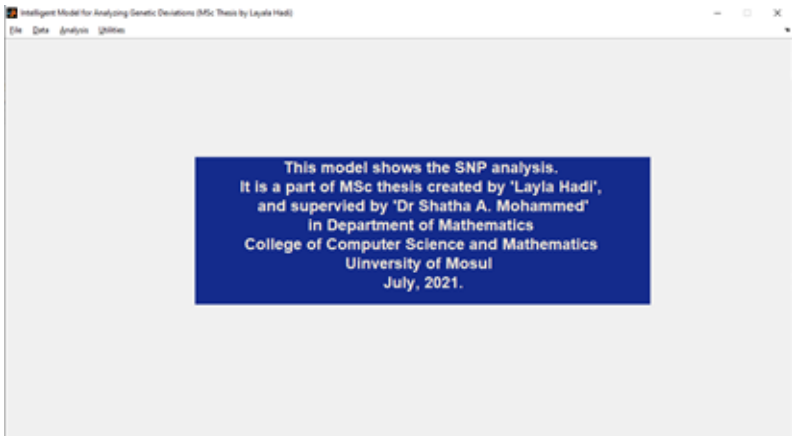
يوضح الجدول (3) أداء الخوارزمية المتقرحة PSO_KPCA على مجموعات بيانات SNP الخاصة بالأمراض المعقدة، إذ تحتوي كل مجموعة منهم على اعداد مختلفة من SNPs. تم مقارنة نتائج الطريقة المقترحة مع Acc وF المحققة باستخدام المعادلتان (1)، (2)، وتقليل الابعاد عن طريق خوارزمية KPCA واختيار الميزات المستندة إلى خوارزمية PSO. كما يمكن أن نرى بوضوح أداء الخوارزمية المقترحة أفضل بكثير وتفوقها على خوارزمية KPCA.

تظهر النتائج التي تم الحصول عليها تفوق الطريقة المقترحة على الخوارزمية الاصلية المقارنة في جميع مجموعات البيانات التي تم اختبارها. تم تحقيق زيادة ملحوظة في متوسط الدقة في جميع مجموعات البيانات المقدمة: على سبيل المثال، الدقة التي تم تحقيقها في مجموعة بيانات CEU للكروموسوم رقم 13، عند استخدام الطريقة المقترحة تم الحصول على افضل دقة تصل إلى 95.7٪. كما أظهرت الخوارزمية المقترحة أداءً ثابتًا عند تطبيقها على مجموعة البيانات المختلفة من خلال إعطاء قيمة عالية لـ Acc وF في جميع البيانات المختبرة.

تثبت النتائج أن النموذج الهجين المقترح يتمتع بأداء قوي وثابت مقارنة بمعظم المصنفات. في مجموعات بيانات، وأظهرت جميع المصنفات المعطاة أداءً رائعًا عند تقييم الخوارزمية المقترحة.

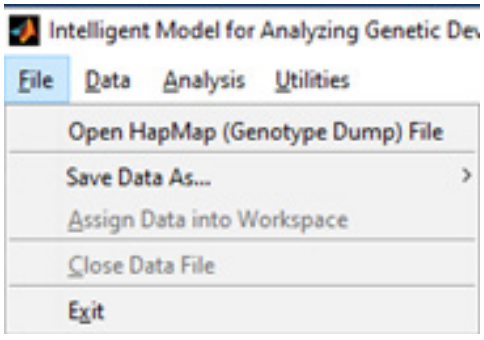
4-4. النموذج المقترح لتحليل الانحرافات الجينية

بالاعتماد على النتائج التي تم الحصول عليها ومناقشتها، تم اقتراح نظام حاسوبي يتطلب تقليل أبعاد البحث للحصول على نسبة عالية في التمييز فضلاً عن سرعة التنفيذ. فتم تقديم مجموعة متنوعة من الدوال لتحليل SNPs في النموذج المقترح، والذي يكون مزوداً بواجهة مستخدم الرسومية GUI، والتي تتكون من اربع قوائم («ملف File»، «بيانات Data»، «تحليل Analysis» و«خدمات Utilities») (شكل (2)).



شكل (2): واجهة التطبيق الرسومية GUI للنموذج المقترح

يتم في قائمة «ملف File» فتح ملف البيانات والذي يكون بامتداد «*.hmp» لقراءة مجموعة البيانات بالتنسيق المحدد بوساطة مشاريع HapMap [30]. شكل (3) يمثل الشاشة الخاصة بالايعاز فتح الملف.

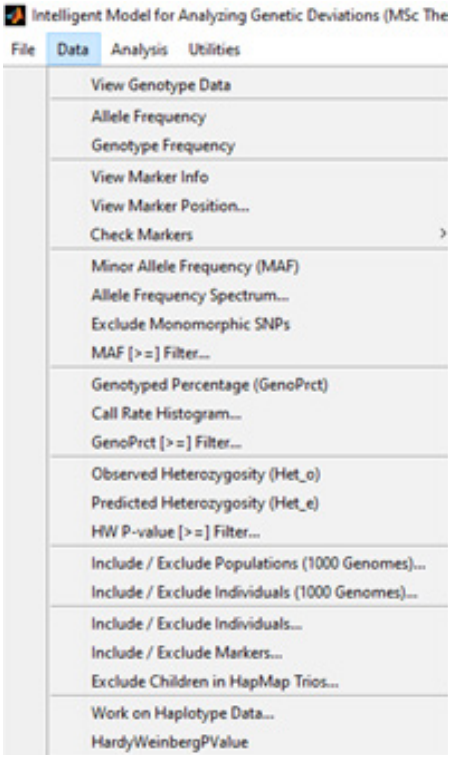


شكل (3): ايعازات قائمة ملف File

بعد قراءة البيانات، يأتي دور التأكد من البيانات واستعراضها، باستخدام قائمة «بيانات Data» (شكل (4)). فبالامكان حساب تغاير الزيجوت Hetero-zygosities الملاحظ والمتوقع، واعداد الأليل الثانوي Minor allele، وقيمة P لاختبار توازن Hardy–Weinberg، وتكرارات الأليل Allele والنمط الجيني Genotype، والاحتمال المركب. كما يستخدم البرنامج المقترح خوارزمية تعظيم التوقع Expectation-Max-imization [31] لتقدير احتمالات الأنماط المنفردة SNPs وحساب إحصائيات اختلال التوازن (LD) Linkage Disequilibrium، مثل D و D' و R، بين أزواج من SNPs. فضلاً عن عرض النتائج ومجموعات البيانات بيانياً. إذ تتضمن بعض الأمثلة مخططًا دائريًا يعرض أليل SNP وتكرارات النمط الجيني لـ11 مجموعات HapMap، ومخطط للمواضع النسبية

لـ SNPs على الكروموسوم، وعرض التركيب الجيني المرئي، كما يقدم مجموعة بيانات خام كاملة للأنماط الجينية للأفراد.

يحسب البرنامج المقترح تنوع النمط المنفرد SNPs [32]، درجة التشابه المنفرد [33]، تماثل الزيجوت في مقطع الكروموسوم [34]، الحد الأدنى لعدد أحداث إعادة التركيب the Minimum number of recombination events () R_m [35]، وتماثل الزيجوت المنفرد الممتد The Extended Haplotype Homozygosity (EHH) [36]، ودرجة النمط المنفرد المتكامل The integrated Haplotype Score



شكل (4): ايعازات قائمة بيانات Data

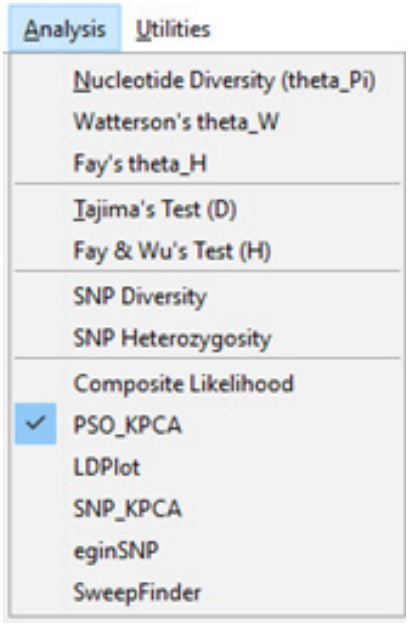
أما بالنسبة لقائمة «تحليل Analysis» فهي القائمة الاساسية (شكل (5))، والتي تعد الهدف الاساسي من اقتراح البرنامج الحالي. إذ تم ادراج الخورازمية المهجنة

(iHS). بالنسبة لـEHH وiHS فهما إحصائيتان قويتان بشكل خاص تم استخدامهما للكشف عن الانتقاء الأخير.

يعتمد EHH على تماثل الزيجوت المنفرد الإحصائي

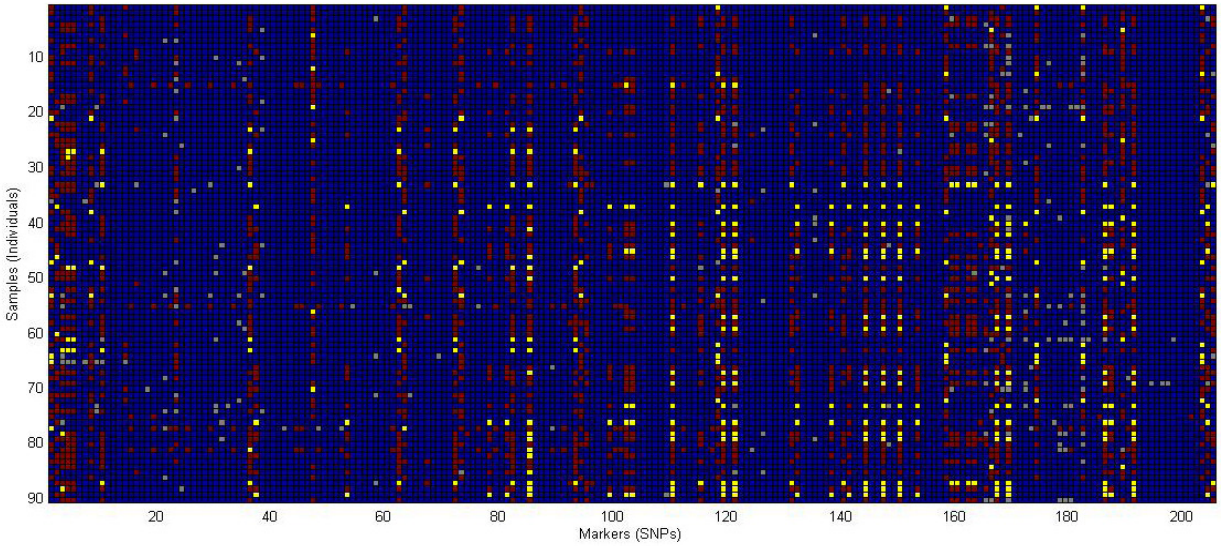
$$HH = \frac{\left(\sum p_i^2 - \frac{1}{n}\right)}{1 - \frac{1}{n}}$$

؛ حيث p_i هو تردد النمط المنفرد النسبي و n حجم العينة. بالنسبة للأنماط المنفردة الأساسية المحددة، EHH تحسب HH بطريقة تدريجية لتحديد كيفية تقسيم LD مع زيادة المسافة إلى منطقة أساسية محددة. يعتمد iHS على المستويات التفاضلية لـ EHH المحيطة بأليل مقارنة بأليل الخلفية في نفس الموضع.



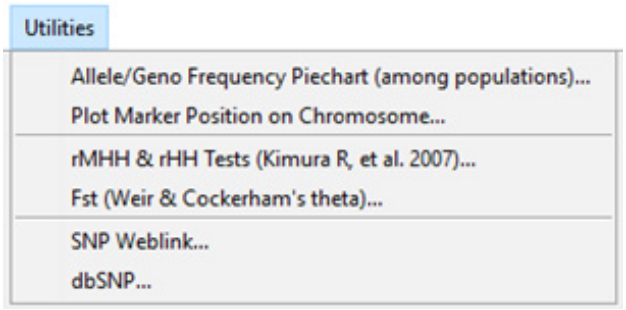
شكل (5): ايعازات قائمة تحليل Analysis

المقترحة لغرض تمييز الـSNPs ضمن مجموعة وتم تسميتها بـ «PSO_KPCA»، انظر (شكل (6)).



شكل (6): ناتج الخوارزمية المهجنة المقترحة PSO_KPCA، والتي تمثل الـ SNPs المصابة بالنسبة للنمط الجيني SNPs لـ Chr13 في المجتمع السكاني CEU

القائمة الاخيرة تضم خدمات Utilities (شكل (7))، تم اقتراحها من قبل الباحثين، فتتوفر فيها رسم شكل لاماكن الـ SNPs في الكروموسم، فضلاً عن اختبارات rMHH، rHH، علاوة على ذلك حساب صيغة Weir لـ Wright's FST، كما يتم الانتقال إلى موقع dbSNP المنضوي تحت الموقع الالكتروني NCBI لاختيار البيانات المطلوبة.



شكل (7): ايعازات قائمة خدمات Utilities

5- الاستنتاجات

اظهرت النتائج التي تم الحصول عليها ان الخوارزمية الذكائية المقترحة PSO_KPCA مهمة في بناء نموذج تصنيف، إذ تعمل عن طريق الحد من عدد ميزات الإدخال في المصنف، بهدف الحصول على نماذج تنبؤية عالية وأقل تعقيدًا من الناحية الحسابية.

في هذه الدراسة تم تهجين خوارزمية تحليل المكونات الاساسية المستندة على النواة KPCA مع خورازمية البحث تحسين سرب الجسيمات PSO، حيث يتم استخدام KPCA كطريقة لتقليل الابعاد للبيانات المطلوبة، واستخدام PSO كطريقة لاختيار الميزات.

وتم تقديم برنامجاً متكاملاً مزوداً بواجهة تطبيق رسومية GUI لتنفيذ الخوارزمية المقترحة، فضلاً عن تطبيق عددًا من الخوارزميات والأساليب والأدوات لإجراء تحليلات استكشافية شاملة للبيانات على مستوى الجينوم. البرنامج المقترح أظهر فائدة Mat-

lab كلغة حسابية علمية قوية ومريحة في علم الوراثة السكانية الجزيئي. إذ يمثل مجموعة أدوات Matlab قوية ومرنة مخصصة لتحليل تعدد الأشكال المنفرد تعدد الأشكال SNPs.

6 - المصادر

1. Kwok, P. Y. (Ed.). (2003). Single nucleotide polymorphisms: methods and protocols (Vol. 212). Springer Science & Business Media.
2. Alzubi, R., Ramzan, N., Alzoubi, H., & Amira, A. (2017). A hybrid feature selection method for complex diseases SNPs. IEEE Access, 6, 1292-1301.
3. Evans, D. T. (2010). A SNP microarray analysis pipeline using machine learning techniques (Doctoral dissertation, Ohio University).
4. Batnyam, N., Gantulga, A., & Oh, S. (2013). An efficient classification for single nucleotide polymorphism (SNP) dataset. In Computer and information science (pp. 171-185). Springer, Heidelberg.
5. Lee, J., Batnyam, N., & Oh, S. (2013). RFS: Efficient feature selection method based on R-value. Computers in biology and medicine, 43(2), 91-99.
6. Liang, J., Yang, S., & Winstanley, A. (2008). Invariant optimal feature selection: A distance discriminant and feature ranking based solution. Pattern Recognition, 41(5), 1429-1439.
7. Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. Machine learning, 53(1), 23-69.
8. Seo, M., & Oh, S. (2012). CBFS: High performance feature selection algorithm based on feature clearness. PloS one, 7(7), e40419.
9. Goudey, B., Abedini, M., Hopper, J. L., Inouye, M., Makalic, E., Schmidt, D. F.,... & Reumann, M. (2015). High performance computing enabling exhaustive analysis of higher order single nucleotide polymorphism interaction in Genome Wide Association Studies. Health information science and systems, 3(1), 1-11.
10. Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L., & Yu, W. (2010). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. The American Journal of Human Genetics, 87(3), 325-340.
11. Yung, L. S., Yang, C., Wan, X., & Yu, W. (2011). GBOOST: a GPU-based tool for detecting gene–gene interactions in genome–wide case control studies. Bioinformatics, 27(9), 1309-1310.
12. Gyenesei, A., Moody, J., Semple, C. A., Haley, C. S., & Wei, W. H. (2012). High-throughput analysis of epistasis in genome-wide association studies with BiForce. Bioinformatics, 28(15), 1957-1964.
13. Zhang, Y., & Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. Nature genetics, 39(9), 1167-1173.
14. Wang, J., Joshi, T., Valliyodan, B., Shi, H., Liang, Y., Nguyen, H. T., & Xu, D. (2015). A Bayesian model for detection of high-order interactions among genetic variants in genome-wide association studies. Bmc Genomics, 16(1), 1-20.
15. Han, B., Chen, X. W., Talebizadeh, Z., & Xu, H. (2012). Genetic studies of complex human diseases: characterizing SNP-disease associations using Bayesian networks. BMC systems biology, 6(3), 1-12.
16. Yang, C., He, Z., Wan, X., Yang, Q., Xue, H., & Yu, W. (2009). SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. Bioinformatics, 25(4), 504-511.
17. James J. Cai, PGEToolbox: A Matlab Toolbox for Population Genetics and Evolution, Journal of Heredity, Vol-

27. Greitans, M., Aristov, V., & Laimina, T. (2012). Application of the Karhunen-Loeve transformation in bio-radiolocation: Breath simulation. Automatic Control and Computer Sciences, 46(1), 18-24.
28. Shang, J., Sun, Y., Li, S., Liu, J. X., Zheng, C. H., & Zhang, J. (2015). An improved opposition-based learning particle swarm optimization for the detection of SNP-SNP interactions. BioMed research international, 2015.
29. Thorisson, G. A., Smith, A. V., Krishnan, L., & Stein, L. D. (2005). The international HapMap project web site. Genome research, 15(11), 1592-1593
30. Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F. L., Yang, H. M., & Duster, T. (2003). The international HapMap project.
31. Moon, T. K. (1996). The expectation-maximization algorithm. IEEE Signal processing magazine, 13(6), 47-60.
32. Depaulis, F., & Veuille, M. (1998). Neutrality tests based on the distribution of haplotypes under an infinite-site model. Molecular biology and evolution, 15(12), 1788-1790.
33. Hamilton, M. B. (2021). Population genetics. John Wiley & Sons.
34. Kijas, J. W., Lenstra, J. A., Hayes, B., Boitard, S., Porto Neto, L. R., San Cristobal, M.,... & International Sheep Genomics Consortium. (2012). Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. PLoS biology, 10(2), e1001258.
35. Avise, J. C. (2000). Phylogeography: the history and formation of species. Harvard university press.
36. Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D.,... & Hurles, M. E. (2006). Global variation in copy number in the human genome. nature, 444(7118), 444-454.

ume 99, Issue 4, July-August 2008, Pages 438–440,
https://doi.org/10.1093/jhered/esm127
18. Miles, L. S., Rivkin, L. R., Johnson, M. T., Munshi□South, J., & Verrelli, B. C. (2019). Gene flow and genetic drift in urban environments. Molecular ecology, 28(18), 4138-4151.
19. Roth, S. C. (2019). What is genomic medicine?. Journal of the Medical Library Association: JMLA, 107(3), 442.
20. O'Rielly, D. D., & Rahman, P. (2021). Clinical and molecular significance of genetic loci associated with psoriatic arthritis. Best Practice & Research Clinical Rheumatology, 101691.
21. Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X.,... & Ming, R. (2018). Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. Nature genetics, 50(11), 1565-1573.
22. Deng, N., Zhou, H., Fan, H., & Yuan, Y. (2017). Single nucleotide polymorphisms and cancer susceptibility. Oncotarget, 8(66), 110635.
23. Bahji, A., Forth, E., Hargreaves, T., & Harkness, K. (2021). Genetic Markers of the Stress Generation Model: A Systematic Review. Psychiatry Research, 114139.
24. Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F. L., Yang, H. M.,... & Duster, T. (2003). The international HapMap project.
25. Borgognone, M. G., Bussi, J., & Hough, G. (2001). Principal component analysis in sensory analysis: covariance or correlation matrix?. Food quality and preference, 12(5-7), 323-326.
26. Lu, H., Meng, Y., Yan, K., & Gao, Z. (2019). Kernel principal component analysis combining rotation forest method for linearly inseparable data. Cognitive Systems Research, 53, 111-122.