



الجامعة المستنصرية/ كلية العلوم

إعداد: م.م مريم جاسم حسن البياتي

المعلومات الحيوية والذكاء الاصطناعي: قراءة البيانات الجينية واستكشافها باستخدام الحوسبة والتعلم الآلي

المحور الأول: المدخل الشامل والتأسيس المفاهيمي

1.1 مفهوم المعلومات الحيوية (Bioinformatics) والتطور التاريخي

تُعرّف المعلومات الحيوية بأنها التخصص العلمي البين-مناهجي (Interdisciplinary) الذي يدمج علوم الأحياء، الحوسبة، الرياضيات، والإحصاء بهدف تخزين البيانات البيولوجية، تنظيمها، تحليلها، وتفسيرها.

مرّ التخصص بثلاث مراحل رئيسية:

١. **مرحلة ما قبل الجينوم (Pre-Genomic Era):** صياغة الخوارزميات الأولى لمقارنة تسلسلات البروتينات (مثل خوارزميات Needleman-Wunsch في بداية السبعينيات).
٢. **مرحلة الجينوم البشري (Human Genome Project - HGP):** تمتد من ١٩٩٠ إلى ٢٠٠٣، وهي المرحلة التي حوّلت العلوم الحيوية إلى علوم كثيفة البيانات بعد قراءة الشفرة الوراثية البشرية كاملة.
٣. **مرحلة البيانات الضخمة والذكاء الاصطناعي (Post-Genomic & AI Era):** التدفق الهائل للبيانات الناتجة عن تقنيات التسلسل الحديثة جعل تحليلات الحوسبة التقليدية غير كافية، مما استدعى إدخال تقنيات التعلم الآلي (Machine Learning) والتعلم العميق (Deep Learning).

1.2 طبيعة البيانات الجينية وخصائصها (Big Data in Genomics)

تتمتع البيانات الجينية بخصائص تجعلها فريدة ومأهولة بالتحديات الحوسبية:

- **الحجم الضخم (Volume):** جينوم بشري واحد يتكون من نحو 3×10^9 زوج من القواعد النيتروجينية، وتخزينه كبيانات خام قد يتجاوز ١٠٠ جيجابايت للعينة الواحدة.

- الأبعاد العالية: ($\text{High-Dimensionality} - \text{\$p \gg n\$}$) غالباً ما يحتوي الجدول الإحصائي للجينوم على ملايين المتغيرات الجينية ($\text{\$p\$}$) لمئات فقط من المرضى أو العينات ($\text{\$n\$}$).
- التعقيد والضوضاء: ($\text{Complexity \& Noise}$) وجود تكرارات جينية واسعة، ومناطق غير مشفرة (Non-coding regions)، بالإضافة إلى الأخطاء العشوائية أثناء عمليات التسلسل.

المحور الثاني: توليد البيانات الجينية والمعالجة الحوسبية الأولية

2.1 تقنيات التسلسل الجيني (DNA Sequencing Technologies)

التقنية	آلية العمل	المزايا	التحديات/القيود
Sanger Sequencing	إنهاء السلسلة بالنوكلีโอتيدات المزدوجة (Chain Termination)	دقة عالية جداً	بطيئة جداً، كلفة عالية، قراءات قصيرة ومحدودة
Next-Generation Sequencing (NGS)	التسلسل الموازي الفائق بالتجميع (Massively Parallel Sequencing)	إنتاجية فائقة، كلفة منخفضة جداً لكل قاعدة	قراءات قصيرة (Short Reads: 100-300 bp) تتطلب تجميعاً معقداً
Long-Read Sequencing (Nanopore/PacBio)	قراءة الجزيئات المفردة بطول كبير	قراءات طويلة جداً تتجاوز ١٠,٠٠٠ قاعدة	نسبة خطأ أعلى نسبياً مقارنةً بـ NGS التقليدي

2.2 أنبوب المعالجة الحاسوبية للبيانات الجينية (Bioinformatics Pipeline)

تمر البيانات الجينية المسلسلة بخطوات متتالية قبل أن تصبح جاهزة للذكاء الاصطناعي:

١. ضبط الجودة: (Quality Control - QC)

- تقييم جودة القواعد باستخدام مؤشر Phred Score (\$Q\$).
- إزالة الوصلات الاصطناعية (Adapters) والقراءات ذات الجودة المنخفضة عبر أدوات مثل FastQC و Trimmomatic.

٢. محاذاة التسلسلات: (Sequence Alignment)

- مطابقة ملايين القراءات القصيرة مع الجينوم المرجعي (Reference Genome) باستخدام خوارزميات تعتمد على التحويلات الرياضية مثل Burrows-Wheeler Transform (BWT) والمستعملة في أدوات BWA و Bowtie2.

٣. تحديد المتغيرات الجينية: (Variant Calling)

- تحديد الفروقات الجينية مثل:
 - تعدد أشكال النوكليوتيدات المفردة (SNPs) تغيير قاعدة واحدة.
 - الإدخال والحذف (Indels) إضافة أو حذف قواعد تؤدي لإزاحة إطار القراءة (Frameshift).
- الاستعانة بأدوات مثل GATK (Genome Analysis Toolkit) لتحديد هذه المتغيرات بدقة إحصائية.

المحور الثالث: تطبيقات الذكاء الاصطناعي والتعلم الآلي في استكشاف الجينات

3.1 التعلم الخاضع للإشراف (Supervised Learning)

يُستخدم عندما تتوفر بيانات تدريب مرفقة بتصنيفات أو قيم معروفة (مثل: مريض vs معافى).

• آلات متجه الدعم: (Support Vector Machines - SVM)

- الآلية: تصنيف البيانات عن طريق البحث عن "المستوى الفائق (Hyperplane)" الذي يحقق أقصى هامش فصل بين المجموعات.
- التطبيق الجيني: التمييز بين الأنسجة السرطانية والأنسجة السليمة بناءً على مستويات التعبير الجيني. (Gene Expression Analysis)

• الغابات العشوائية: (Random Forests)

- الآلية: دمج مئات من أشجار القرار لحساب أهمية المتغيرات وتقليل ظاهرة الإفراط في التلاؤم. (Overfitting)
- التطبيق الجيني: تحديد الجينات الأكثر تأثيراً في تطور الأمراض المعقدة مثل السكري أو ألزهايمر. (Feature Importance Selection)

3.2 التعلم غير الخاضع للإشراف (Unsupervised Learning)

يهدف لاكتشاف الأنماط والتجمعات المخفية داخل البيانات دون وجود تصنيفات مسبقة.

• التحليل العنقودي: (Clustering)

- مثل خوارزميات K-Means و Hierarchical Clustering.
- التطبيق الجيني: اكتشاف أنماط فرعية جديدة للأنواع السرطانية (Cancer Subtypes) غير المكتشفة سريرياً.

• تقليل الأبعاد: (Dimensionality Reduction)

- مثل (PCA) تحليل المكونات الرئيسية) و t-SNE و UMAP.
- التطبيق الجيني: ضغط ملايين المتغيرات الجينية لتمثيل العينات في بُعدين أو ثلاثة أبعاد لرؤية العلاقات بين الجماعات والسكان.

3.3 التعلم العميق ونمذجة التسلسلات كـ "لغات بيولوجية-Deep Learning & Bio-NLP"

تتعامل تقنيات التعلم العميق الحديثة مع تسلسل DNA كأنه لغة طبيعية، حيث القواعد النيروجينية هي الحروف، والجينات هي الكلمات:

• الشبكات العصبية الالتفافية: (CNNs)

- التطبيق: مسح تسلسلات DNA باكتشاف الأنماط الموضعية القصيرة (Motifs)، وتحديد مواقع ارتباط البروتينات بالحامض النووي (مثل نموذج DeepSEA).

• نماذج المحولات (Transformers) ونمذجة اللغة الجينية:

- مثل نموذج DNABERT ونماذج Nucleotide Transformer.
- الآلية: استخدام آليات الاهتمام الذاتي (Self-Attention) لفهم العلاقات بعيدة المدى بين القواعد النيروجينية والتنبؤ بمناطق المحفزات (Promoters) والمُعززات (Enhancers).

• التنبؤ ببنية البروتينات: (AlphaFold)

- إنجاز شركة DeepMind عبر نموذج AlphaFold 2 & 3، الذي حل مشكلة طي البروتين (Protein Folding Problem) المستمرة منذ ٥٠ عاماً عبر التنبؤ بالشكل الثلاثي الأبعاد بدقة الذرة بناءً على تسلسل الأحماض الأمينية فقط.

المحور الرابع: التطبيقات السريرية والطب الدقيق والتحديات

4.1 الطب الدقيق والعلاج الشخصي (Precision Medicine)

١. الصيدلة الجينية: (Pharmacogenomics) تحليل الجينوم للتنبؤ باستجابة المريض لدواء معين وتحديد الجرعة الآمنة وتجنب التسمم الدوائي.
٢. الخزعة السائلة: (Liquid Biopsy) استخدام الذكاء الاصطناعي لرصد أجزاء الحمض النووي الورمي المتداول في الدم (ctDNA) ، مما يسمح بالكشف المبكر عن السرطان قبل ظهوره بالأشعة.

4.2 التحديات العلمية والأخلاقية

١. مشكلة القابلية للتفسير: (Explainability / Black Box) صعوبة فهم كيفية اتخاذ نماذج التعلم العميق لقرارات تشخيصية، وهو ما يتطلب تطوير مجال الذكاء الاصطناعي القابل للتفسير. (XAI)
٢. التحيز في بيانات التدريب: (Data Bias) معظم أبحاث الجينوم الحالية تعتمد على عينات من أصول أوروبية، مما يجعل النماذج أقل دقة عند تطبيقها على شعوب أخرى.
٣. أمان البيانات وخصوصية المريض: حماية البيانات الجينية الحساسة من الاختراق أو الاستغلال غير الأخلاقي.

المراجع (References)

1. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). *Highly accurate protein structure prediction with AlphaFold*. *Nature*, 596(7873), 583–589.
2. Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Batzoglou, S. (2019). *A primer on deep learning in genomics*. *Nature Genetics*, 51(1), 12–18.
3. Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). *DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome sequence analysis*. *Bioinformatics*, 37(15), 2112–2120.
4. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. خاص بأسس خوارزميات التعلم العميق المطبقة في التحليل الحيوية.