HW10_111652041 報告

在我上個學期修習完的計算分子生物學中，我理解到其實還有許多事情我們是未知的。隨著基因定序技術與生物資訊的快速發展，人類已經能夠完全讀取人類約三十億個鹼基的序列，但完整找出人類哪些字串為我們的基因，仍是一個尚未真正解決的核心問題。目前的演算法主要依賴序列比對、開放閱讀框的搜尋或表達證據來推測基因位置，但仍有大量未知基因、非典型轉錄產物與功能性非編碼序列未被精準辨識。因此，我認為二十年後的人工智慧可能具備一項重要能力！也就是能夠自動解讀整個人類基因組，完整且準確地找出所有基因、轉錄變異與調控功能，讓人類第一次真正掌握基因組中每段序列的角色。

我認為這項能力具有重大科學與醫學意義。若 AI 能辨識每一個基因及其剪接形式、調控區域與細胞類型差異，就能大幅提升對疾病、罕見突變、癌症基因體異常以及基因治療標靶的理解。現今的人類基因組計畫已成功讀出所有鹼基續對，但基因「在哪裡、做什麼、我們該如何調控」仍沒有完整答案。也就是，序列已經存在，但「解讀」對於我們仍是最大的挑戰。如果未來的 AI 能自動進行基因註釋與功能預測，它將不只是工具，而會成為可以主動做科學推論的研究者。

為了讓 AI 擁有這樣的能力，需要結合多種機器學習方法，而非單一模型。首先，由於大量基因尚無標籤，也不存在明確的正確答案，因此非監督式學習是必要的。AI 必須先從全基因組與轉錄體資料中自行找出共同的序列特徵、啟動子語法、剪接訊號與表現群組，而不是只依賴人類輸入的資料。其次，深度學習也扮演關鍵角色。基因序列並非單純的文字，而是一種高度複雜的生物語言，必須依靠卷積神經網路、Transformer 或圖神經網路來解析序列語法與三維染色體折疊關係。最後，監督式學習仍然必要，它可以利用現有已知基因作為訓練資料，讓模型更精確地判斷未知序列是否具有基因功能。因此，未來的最佳策略我猜很可能是非監督式學習、深度神經網路與監督微調的混合架構！

若要研究這項能力，第一步可能可以設計一個簡化的模型化問題：讓 AI 僅利用人類基因組序列與 RNA-seq 的表現資料，自動找出尚未註釋的基因，並預測其剪接位點與可能的轉錄產物。模型的成功可以透過多種方式評估，例如找到的序列是否具有表達證據、是否能正確預測轉錄起點、是否能跨不同細胞或資料庫重複驗證，甚至是否能預測突變造成的功能變化。如果 AI 能夠發現科學界尚未標記的功能性基因，並提供可驗證的預測，這代表模型具有真正的生物意義了！

現代生物資訊與基因體學已經奠定基礎，但距離我們能夠自動找出人類所有基因仍有非常大差距。二十年後的 AI 若能整合序列、轉錄體、表觀遺傳與三維基因組架構，並以自主的方式分析、推論並提出假設，人類將能第一次真正完成基因清單的全貌。這將使疾病研究、精準醫療以及基因治療獲得關鍵的大突破，也可能改變計算分子生物學的研究方式，從人類解讀資料，轉向 AI 主動發現生物規律。