生物資訊與人工智慧的發展現況與展望

Bioinformatics and Artificial Intelligence: Current Progress and the Future

吳瑞裕

摘要

分子生物學上的研究越來越成爲由逐漸累積的資料庫所推動與掌控，其中主要是因爲有了大規模功能性基因體(genomics)和蛋白質體(proteomics)實驗的相繼出現，例如科學家們可以全面地測量基因的表達。這些經由染色體中基因密碼所產生數以萬計的蛋白質，提供了非常豐富的資訊。所以，在生物資訊(bioinformatics)中的挑戰將是如何整合資料庫去連接各式各樣的資訊，並且進行大規模研究來集體分析許多不同的資料集。這種方法代表我們已從傳統只探討單一基因的生物學轉移另一個全新的領域，而且它經常包含了對於某些於特徵點進行統計分析；例如，蛋白質摺疊，質功能，蛋白質交互作用，隱性基因(pseudogenes)，或蛋白質的定位化(localization)。除此之外，在眾多已知的生物資料下，電腦人工智慧(computing intelligence)技術很明顯的可以用來探索生物作用的趨勢和模式。在這篇文章裡，作者摘要出一些電腦人工智慧應用在基因體及蛋白質體上的例子：如，使用叢集法來統整 microarray 表現的資料，利用向量機器(vector machine)預測蛋白質的作用，以及利用 Bayesian networks 預測蛋白質在細胞中的位置等。

Abstract

The research in molecular biology is becoming ever more database driven, motivated nowadays, in part, by the introduction of large-scale functional genomics and proteomics experiments such as those comprehensively measuring gene expression. These provide abundant information on thousands of proteins which encoded by genome. Consequently, a challenge in bioinformatics is integrating databases to connect this unrelated information as well as performing large-scale studies to collectively analyze many different data sets. This approach represents an example shift away from traditional biology, and it often involves statistical analyses focusing on the occurrence of particular features (e.g., protein functions, foldings, interactions, pseudogenes, or localization) in genomics and proteomics. In addition, the application of computing intelligence techniques can be used to discover trends and patterns in the primary data. In this article, author gave several examples of these techniques in a genomic and proteomic context: clustering methods to organize microarray expression data, support vector machines to predict protein function, Bayesian networks to predict sub-cellular localization, and decision trees to optimize target selection

for high-throughput proteomics.