

利用雲端運算解決蛋白質摺疊

Using Cloud Computing To Solve Protein Folding Problem

指導教授：謝孫源

專題成員：林威宏

開發工具：Hadoop、JAVAFX

測試環境：Ubuntu 14.04

一、簡介

蛋白質摺疊問題 (Protein Folding Problem) 被列為21世紀的生物物理學上的一個重要未解問題。蛋白質的基本單位為胺基酸，而蛋白質的一級結構指的就是其氨基酸序列，蛋白質會由所含的氨基酸殘基的親水性、疏水性、帶正電、帶負電、凡德瓦力等特性通過殘基間的相互作用而摺疊成一立體的三級結構。要解決的是從給定的氨基酸序列，預測它在三維空間下可能產生的折疊結構。雖然蛋白質可在短時間中從一級結構摺疊至立體結構，研究者卻無法在短時間中從氨基酸序列計算出蛋白質結構，甚至無法得到準確的三維結構。

在這個問題中，影響最大的因素為疏水力。而胺基酸被分為疏水性 (H) 與極性 (P) 兩類。當兩個序列中不連續的H胺基酸，在晶格結構中相鄰時，兩者間會產生疏水力，導致cost降低。我們的目標便是找出cost最小的折疊結構。

我們將胺基酸HP序列寫成串列的形式 (EX: HPPHP...)，表示它們在晶格中是有鍵結的。程式採用分支定界 (Branch & Bound) 的方式，取串列前面的一段，計算出目前可能產生的所有結構和分別的cost，並從中挑出cost 最小者。之後便以上述方式，持續到串列取完為止，便能挑出擁有最小cost 的折疊結構、摺疊成為某一具有最低能量狀態的穩定結構。

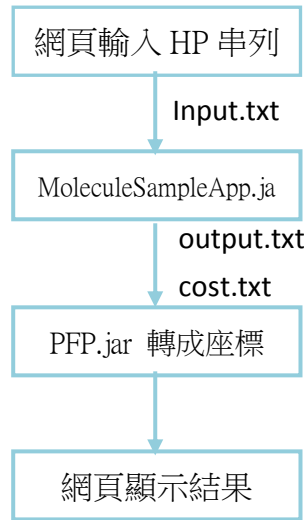
● 平行運算:

為了縮短Protein Folding Problem 中min cost 的計算時間，我們利用Hadoop的平行運算架構，將問題分散到不同機器上同時計算，以加快解答速度。一開始先取HP 序列中的一小段，排列出可能產生的結構，將這些結構分派 (map)到指定的電腦上。每台電腦再取序列的一小段來計算，分別找出各自的min cost。將每台電腦找出擁有min cost結構的結果回收(reduce)至一台指定的電腦上，比較這些結果，找到其中有最小min cost的結構，再將它分派出去。重複這個步驟，最後便會找出整個序列的min cost結構。

- JAVA FX

而為了使得出最佳結構的結果視覺化，將運用JAVA FX的技術以及內建的模型物件完成。先模擬出三維空間，再將所給HP座標資料一次一點對應到座標軸上。每放上一點，便與前一點算出中點，即兩點間之鍵結位置。最後判斷每個H之間是否需要加上cost，以完成HP模型呈現。

- 架構圖



網頁的流程是先輸入 HP 串列，經後端處理記錄在文字檔 input.txt 中。呼叫 PFP.jar 算出 input.txt 中串列的最佳折疊方式，並將其座標點及 min cost 分別輸出到 output.txt 和 cost.txt 中。MoleculeSampleApp.jar 依照 output.txt 中的座標點建出 3D 模型並嵌入網頁中，完成蛋白質最佳折疊結構預測與視覺化。

二、測試結果

The screenshot shows the 'Protein Folding Prediction' web application. At the top, there are navigation links: 'home', 'about', 'document', 'Q&A', and 'reference'. The main content area is divided into three sections: '輸入說明' (Input Instructions), '模型操作' (Model Operations), and '注意事項' (Notes). The '輸入說明' section lists five steps for using the application. The '模型操作' section lists five keyboard shortcuts for interacting with the 3D model. The '注意事項' section lists three important notes. In the center, a 3D molecular model is displayed on a black background with a grid. To the right of the model is a text area showing the output coordinates and the minimum cost. At the bottom, there is an input field for the 'HP sequence' and two buttons: 'Submit' and 'Reset'.

輸入說明：

1. 在"HP sequence"輸入HP序列。ex:HPPH...
2. submit鍵 送出序列；reset鍵清除輸入。
3. 待運算完成後，中間顯示預測的3D模型，綠點是H，白點是P，黃色虛線是cost。
4. 右邊分別顯示模型的座標點及min cost。

模型操作：

1. 左鍵拖曳，調整視角
2. 右鍵左右拖曳，調整視距
3. Ctrl+X 隱藏/顯示坐標軸
4. Ctrl+V 隱藏/顯示HP模型
5. Ctrl+Z 清除所有調整

注意事項：

1. 需要安裝JAVA 8_20以上才能顯示結果。
2. 模型無法顯示，請參考Q&A。
3. input長度限制：20。

HP sequence : Submit Reset

min cost: -37