用 CUDA 進行局部序列比對

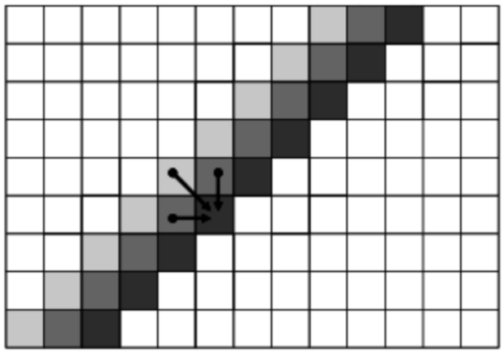
Semi Global Sequence Alignment with CUDA

指導教授：賀保羅
專題成員：吳信葆
開發工具：GCC、CUDA、python
測試環境：Ubuntu

一、簡介：

人工智慧之所以能夠普及，是因為電腦的計算能力大幅提升，加上 GPU 的平行運算能力，這不經讓人開始思考，對於一些因為計算上較為耗時的演算法，用 GPU 加速之後，是否就可以讓該演算法重新地被人認識。

所以本次專題，我將會拿序列比對演算法進行優化，這個演算法屬於 Dynamic Programing 的一種，從 Data Dependency 角度出發，可以發現非常適合平行化，然而 DP 的可平行化運算，是大量的小任務，並且要等到所有小任務執行完之後，才可以跑下一批，這就導致如果使用 CPU 去跑這些任務需要頻繁的去增設 barrier，並且每一個 CPU thread 的成本也很高，不論是建立或是讓更多 CPU thread 溝通的 latency。



但是這樣子的情況卻非常適合 GPU，總共有兩點原因。

第一點，GPU 的 thread 很多。

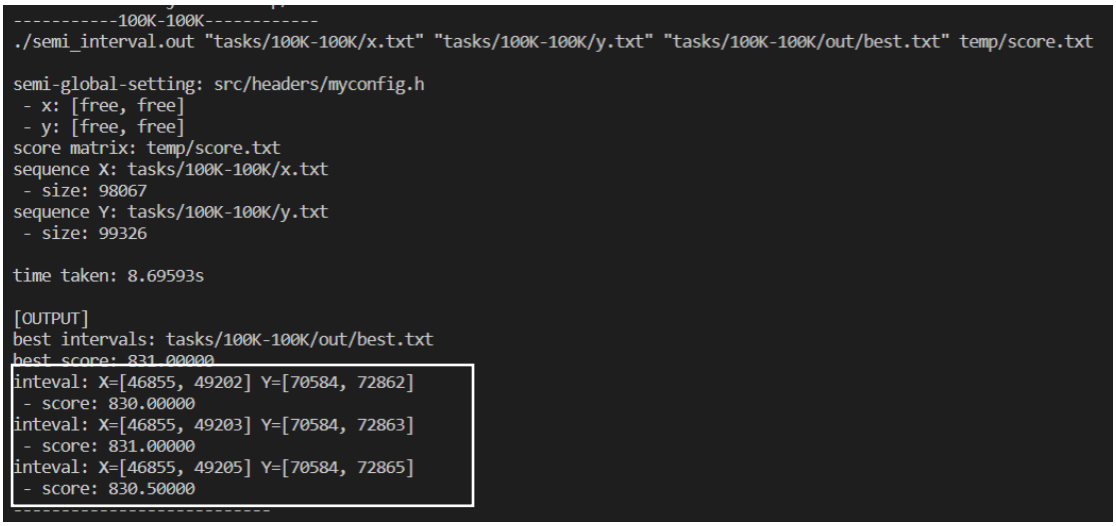
第二點，呼叫 CUDA kernel (threads) 的 latency 很低，這就表示我們每次都能一次呼叫大量的 thread 去處理每批任務，每批任務重新呼叫。

並且我還在這樣子的基礎下，實作兩個新的演算法

第一，Semi Global Interval，因為如果只是單純的全局 (Global) 比對兩條序列，分數會很低，大部分時候是沒有意義的。通常會有幾段序列區間彼此很像，也就是比對分數很高，所以這個演算法會把這些區段截取出來。

第二，Alignment，對截取出來的區段進行 Global 序列比對，達成專題題目承諾的要求。

二、測試結果：



上圖是對兩條長度為 10 萬的 DNA 進行 Semi Global Interval 程式的結果，輸出了三條比對區間，第一條是 X 序列在 46855 到 49202 的區間，Y 序列在 70584 及 72862 之間，分數是 830.0 分，與最高分在 3 分差距以內，故被截取出來。

接著對第一對 DNA 區段進行 Alignment

```
X TATGATATGATGTGGCTGTGTCCCCACCCAAATCTCATCTTGAATTCTAC
Y TATGATATGGTTTGGCTGTATCCCAACTCAAATCCCATTATGAATTGTAT

X TTCCCATAAATCCCCACGTGTCATGGGAGGGACCCAGTGGGAGATAATCA
Y CTCCCATAATTCCC-ACATGTTGTGGGAGGGACCTGGTGGGACATAATTG

X AATCATGGGGGCGGTTACCCCCATGCTGCTATTCTTGTGATAGTAAGTGA
Y AATCCAGGGAGTGGTTTCCCCCGTACTG---GGCTTGTGGTAGTGAAT-A

X GTTCTCATGAGATCTGATGGTTTTATAAGGGGCTTTTCCTCCCTTTACTT
Y TATCTCACAAGATCTGATGATTTTACAAGGGG--TTTCC-CCTTTCACTT

X GGCACT----TCTC-CTTC-CTGCCATCATGTGAATAAGGACACGTTTGT
Y GGCACTCCTTTCTCTCTTGTCCGCCACCATGT----AAG-ACA--TGCCT

X TTCCCCTTTCACTAAGATTGTAAGCTTCCTGAGGCCTCCCCAGCCATGTG
Y TTCCCCTT---CTGCCAT-G-A--CT--CTGAGGCCTCCCCAGCCATGTG
```

得到最終比對結果，奇數行是 X 偶數行是 Y，因為比對太長所以換行。
使用 CUDA 後的效能評測及分析會在專題報告中展示。