Design Group
D17-001

適用於次世代基因定序之高速資料處理器

A Fully Integrated Data Processor for Next-Generation Sequencing

隊伍名稱 沙漠中的讚美
Praise in the Desert

隊　　長 吳易忠 / 臺灣大學電子工程學研究所

指導教授
楊家驤
臺灣大學電機工程學系

研究領域
生醫訊號處理器、基頻通訊積體電路、機器學習處理晶片設計。

美國加州大學洛杉磯分校電機博士，現為臺灣大學電機工程學系副教授，實驗室致力於開發低功耗之客制化晶片以提升資料處理速度與能量效率。曾任教於交通大學電子工程學系。

指導教授
洪瑞鴻
交通大學資訊工程學系

研究領域
次世代定序演算法、生物資訊分析、複雜系統與機器學習。

美國波士頓大學博士，現為交通大學資訊工程學系副教授，實驗室致力於開發更高效的演算法與計算模型來了解基因體的複雜問題。

作品摘要

【研究動機】

基因工程相關研究是近年來生物與醫療領域不可或缺的一部分。隨著醫療產業的進步與發達，人們對於快速的基因疾病偵測和篩檢的需求也逐漸增加。次世代基因定序（Next-Generation Sequencing，NGS）可透過大量平行的方式產生大量的短片段，進一步得到人體內的含氮鹼基（A、C、G、T）的排列順序。定序完成的基因序列則需要經過後續的資料分析步驟方能用於疾病檢測與診斷。由於人體內的基因序列長度超過三十億個含氮鹼基對，以現今透過軟體進行的資料分析與處理時間仍然需要消耗大量的計算時間。本設計提出了第一顆次世代基因定序資料分析晶片，實現了 sBWT 演算法並完全的整合了基因分析演算法當中最複雜的後綴字串排序、建表與搜尋的部分。並且採用了大量平行化的硬體排序電路以達到高通量即時排序，可將次世代基因資料分析的後綴字串排序與建表所需時間降低至十分鐘之內，並同時達到高速短片段搜尋的功能。

【系統架構與硬體突破性】

1. 傳統之資料分析演算法採用 hash table 建表並進行搜尋，佔用過大記憶體容量。而一般 FM-index 之 memory footprint 過大，因此較難使用硬體實作。本作品使用 sBWT 演算法，可建立 k-ordered FM-index 資料結構，大量降低排序時的運算複雜度，並可將該資料結構所使用之記憶體空間根據不同系統需求做調整，達到硬體完全可實現的目標（1GB 以內）。
2. 於建表當中所需使用之排序演算法，原本由軟體實現之複雜度過大。本作品透過大量平行之硬體排序單元，大幅降低排序演算法中的運算延遲。並透過桶排序演算法與升 / 降取樣之設計，可大量降低排序的運算複雜度、平均運算負載，有效提升建表速度。
3. 傳統軟體演算法於字串搜尋中無法達到平行比對的效果。本作品使用了多個平行比對模組，可快速還原 k-ordered FM-index 之中被降取樣掉的資料，相較於無平行處理之設計節省了 98% 的運算時間，不僅可節省記憶體空間、亦可達到快速搜尋之目的。
4. 本作品亦設計了參數分析模組，可動態調整不同長度的基因序列之最佳運算參數。於記憶體需求與搜尋速度上找出最佳設計參數，以達到最佳硬體效能。

【實作成果】

本作品提出之次世代基因定序資料處理器以台積電 40nm CMOS 製程實現，晶片面積為 7.84mm^2，邏輯閘數為 7.4M，操作頻率為 200MHz。操作在 0.9V，晶片功耗為 135mW，能量效率為 3.7×104。相較於傳統高階 CPU 與 GPU，有數個量級（8000×）以上的增益。

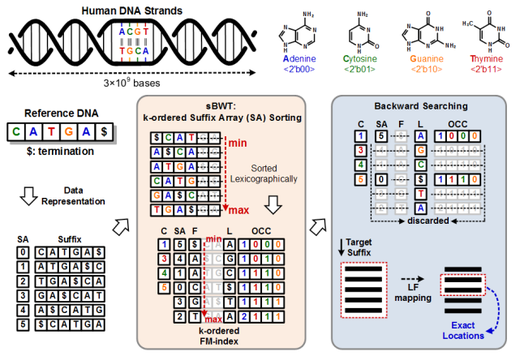

圖 1. NGS

Abstract

【Motivation】

Genetic engineering is an indispensable field for biomedical researches nowadays. Next generation sequencing （NGS） enables high-throughput sequencing, in which short DNA fragments can be sequenced in a massively parallel fashion. However, the essential algorithm behind the succeeding NGS data analysis, DNA mapping, is still excessively time consuming. Based on the memory-efficient sBWT algorithm, this work is the first integrated NGS data processor for the practical NGS data analysis. The k-ordered FM-index used in the sBWT algorithm is leveraged to improve storage capacity and reduce hardware complexity. The proposed NGS data processor realizes the sBWT algorithm through novel Bucket Sorting, Suffix Grouping, and Suffix Sorting circuits. The most complicated part, SA sorting, can now be finished within ten minutes.

【System model and breakthrough】

The conventional algorithms uses hash table to perform DNA mapping, but too much memory space is required. The FM-index based induces large memory footprints and thus is not a proper way for hardware implementation. In this work, sBWT algorithm is utilized to construct the k-ordered FM-index, which can greatly reduce the hardware complexity. In addition, it can also scale down the memory space to meet the system requirement and bring a feasible solution for hardware realization within 1 Gigabyte.

The software implementation introduces heavy computation loads for sorting. In this work, massively parallel insertion sorting elements are utilized to reduce the sorting latency. By exploiting bucket sort with up/down sampling scheme, the computation complexity and workload are greatly reduced, thus enhancing the throughput for FM-index construction.

Conventional software algorithm is not able to execute in parallel when performing target string matching. In this work, several parallel matching modules are used to recover the down-sampled data in FM-index. Compared to the baseline design, the searching latency is saved by 98%.

The key design parameters are analyzed and selected dynamically to reach the optimal performances for different lengths of DNA sequences. Parameters are adaptively derived to achieve maximal searching throughput under given memory spaces and DNA sequence.

【Experimental Results and Conclusion】

Fabricated in the 40nm process, the chip integrates 7.4M gates in 7.84mm^2 area. The power dissipation is 135mW at 200MHz from a 0.9V supply voltage. This work is the first dedicated NGS data processor that implements both SA sorting and backward searching on silicon. It achieves more than 8,000x higher energy efficiency to high-end GPU solutions.