Generalized Pattern Matching Applied to Genetic Analysis

Bing Ni

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
Computer Science and Engineering

The Chinese University of Hong Kong
Mar 2011

Abstract of thesis entitled:

Generalized Pattern Matching Applied to Genetic Analysis

Submitted by Bing Ni
for the degree of Doctor of Philosophy
at The Chinese University of Hong Kong in Mar 2011

Approximate pattern matching problem is, given a reference sequence T, a pattern (query) Q, and a maximum allowed error e, to find all the substrings in the reference, such that the edit distance between the substrings and the pattern is smaller than or equal to the maximum allowed error. Though it is a well-studied problem in Computer Science, it gains a resurrection in Bioinformatics in recent years, largely due to the emergence of the next-generation high-throughput sequencing technologies. This thesis contributes in a novel generalized pattern matching framework, and applies it to solve pattern matching problems in general and alternative splicing detection (AS) in particular. AS is to map a large amount of next-generation sequencing short reads data to a reference human genome, which is the first and an important step in analyzing the sequenced data for further Biological analysis. The four parts of my research are as follows.

In the first part of my research work, we propose a novel deterministic pattern matching algorithm which applies `Agrep`, a well-known bit-parallel matching algorithm, to a truncated suffix array. Due to the linear cost of `Agrep`, the cost of our approach is linear to the number of characters processed in the truncated suffix array. We analyze the matching cost theoretically, and obtain empirical costs from experiments. We carry

out experiments using both synthetic and real DNA sequence data (queries) and search them in Chromosome-X of a reference human genome. The experimental results show that our approach achieves a speed-up of several magnitudes over standard `Agrep` algorithm.

In the second part, we define a novel generalized pattern (query) and a framework of generalized pattern matching, for which we propose a heuristic matching algorithm. Simply speaking, a generalized pattern is $Q_1G_1Q_2...Q_{c-1}G_{c-1}Q_c$, which consists of several substrings Q_i and gaps G_i occurring in-between two substrings. The prototypes of the generalized pattern come from several real Biological problems that can all be modeled as generalized pattern matching problems. Based on a well-known seeding-and-extending heuristic, we propose a dual-seeding strategy, with which we solve the matching problem effectively and efficiently. We also develop a specialized matching tool called `Gpattern-match`. We carry out experiments using 10,000 generalized patterns and search them in a reference human genome (hg18). Over 98.74% of them can be recovered from the reference. It takes 1-2 seconds on average to recover a pattern, and memory peak goes to a little bit more than 1G.

In the third part, a natural extension of the second part, we model a real biological problem, alternative splicing detection, into a generalized pattern matching problem, and solve it using a proposed bi-directional seeding-and-extending algorithm. Different from all the other tools which depend on third-party tools, our mapping tool, `ABMapper`, is not only stand-alone but performs unbiased alignments. We carry out experiments using 427,786 real next-generation sequencing short reads data (queries) and align them back to a reference human genome (hg18). `ABMapper` achieves 98.92% accuracy and 98.17% recall rate, and is much better than the other state-of-the-art tools: `SpliceMap` achieves 94.28% accuracy and 78.13% recall rate;

while TopHat 88.99% accuracy and 76.33% recall rate. When the seed length is set to 12 in ABMapper, the whole searching and alignment process takes about 20 minutes, and memory peak goes to a little bit more than 2G.

In the fourth part, we focus on the seeding strategies for alternative splicing detection. We review the history of seeding-and-extending (SAE), and assess both theoretically and empirically the seeding strategies adopted in existing splicing detection tools, including Bowtie's heuristic and ABMapper's exact seedings, against the novel complementary quad-seeding strategy we proposed and the corresponding novel splice detection tool called CS4splice, which can handle inexact seeding (with errors) and all 3 types of errors including mismatch (substitution), insertion, and deletion. We carry out experiments using short reads (queries) of length 105bp comprised of several data sets consisting of various levels of errors, and align them back to a reference human genome (hg18). On average, CS4splice can align 88. 44% (recall rate) of 427,786 short reads perfectly back to the reference; while the other existing tools achieve much smaller recall rates: SpliceMap 48.72%, MapSplice 58.41%, and ABMapper 51.39%. The accuracies of CS4splice are also the highest or very close to the highest in all the experiments carried out. But due to the complementary quad-seeding that CS4splice use, it takes more computational resources, about twice (or more) of the other alternative splicing detection tools, which we think is practicable and worthy.

論文摘要

通用性模式匹配在基因序列分析中的應用

模糊模式匹配問題(approximate pattern matching,以下簡稱**模式匹配**)可以做以下的定義：給定一條參考序列 T，一條查詢序列 Q，和一個最大的允許錯誤 e，我們要尋找參考序列中所有的子序列，使得這些子序列和查詢序列的距離小於或者等於最大的允許錯誤。儘管這個模式匹配問題在計算機科學領域已經得到了深入而徹底的研究，近年來它在信息生物學領域中被重新認識和研究，而這很大程度上是得益於下一代高通量測序(next-generation high-throughput sequencing)技術的出現。本論文著力於提出一種新型的通用性序列的模式匹配模型，並將其應用於解決廣泛性的模式匹配問題，尤其是基因生物學中的選擇性剪接問題(alternative splicing detection,以下簡稱**剪接問題**)。剪接問題是要將下一代測序技術所得的大量短序列數據通過序列比對(sequence alignment)對應到一個參考序列(譬如人類基因)中。這是基因序列分析的第一步，也是相當重要的一個步驟，因為生物學者要基於剪接後生成的結果進行進一步的生物學分析。本論文有四個主要的部分，下面將一一闡述。

在本論文的第一部分中，我們提出一種全新的確定性(deterministic)模式匹配算法：把一個知名的基於位並行處理的匹配算法，Agrep，應用在後綴樹(suffix tree)上。因為 Agrep 的計算複雜度是線性的，我們所提出匹配算法的耗費(cost)也就同它所遍歷的後綴樹所包含的字符個數成線性正比例關係。我們對此匹配算法的耗費進行了理論分析，並且構建真實的後綴樹來計算它的實際耗費。我們在實驗中使用了人造的和真實的乙型肝炎病毒基因(HBV)基因序列(查詢序列)，並在人類基因 X 染色體的參考序列中執行搜索。實驗結果表明，我們提出的方法相比較於標準的 Agrep 算法要快幾個數量級。

在第二部分，我們定義了一個全新的通用性序列(generalized pattern)和通用性序列匹配的模型，並為之提出一個非確定性(heuristic)的算法。簡單來講，一個通用性序列就是 $Q_1G_1Q_2...Q_{c-1}G_{c-1}Q_c$：它包含幾個子序列 Q_i，以及臨近子序列之間的空隙 G_i。這個通用性序列的原型來自於生物學應用中的多個真實問題，而這些問題可以被建模成為通用性序列的模式匹配問題。基於一個知名的種子搜索和延伸比對的策略(seeding-and-extending, 以下簡稱**種子策略**)，我們提出了一個新的雙種子(dual-seeding)搜索的策略，準確而高效地解決了通用性序列的匹配問題。基於此策

略，我們還開發了一個專門解決通用性序列匹配的軟件工具，叫 Gpattern-match。我們在實驗中使用了 10,000 個通用性序列，並在一個人類全基因的參考序列中執行搜索。實驗結果表明，Gpattern-match 能夠成功搜索到 98.74%的查詢序列；平均只需要 1-2 秒就能夠執行一個查詢，而內存的最高值只有 1G 多一點。

在第三部份，這也是第二部分工作的一個自然延伸，我們把一個真實的生物學問題，剪接問題，建模成一個通用性序列的匹配問題。其中，我們提出一個新的基於種子策略的雙向延伸方法，並且開發了一款叫 ABMapper 的軟件工具。以往那些能夠解決剪接問題的軟件工具都要依賴於第三方軟件，ABMapper 則是完全獨立的；不僅如此，它的搜索和比對過程也是全面的。我們在實驗中使用了 427,786 條短序列(下一代基因序列)，並將之比對到一個人類全基因的參考序列上。ABMapper 有 98.92%的準確率和 98.17%的召回率，相比於其他最新的軟件工作要好得多；譬如，SpliceMap 有 94.28%的準確率和 78.13%的召回率，TopHat 有 88.99%的準確率和 76.33%的召回率。如果 ABMapper 將種子長度設定在 12，那麼整個搜索和比對的過程只需要耗費大概 20 分鐘，而內存的最高值只有 2G 多一點。

在第四部分，我們著重於研究剪接問題中的種子策略：回顧了種子策略的起源，並深入評價和比較了現有的剪接軟件工具所用的種子策略，包括 Bowtie 的非確定性種子和 ABMapper 的不容許錯誤的種子策略，和我們提出的一個全新的互補性 4 種子(complementary quad-seeding)策略，以及基於此而開發的一個新的軟件工具 CS4splice。CS4splice 不僅可以容許種子裡的錯誤，還可以解決在搜索和比對過程中可能發生的所有 3 種錯誤，包括替換 (substitution / mismatch)，插入(insertion)，和刪除(deletion)。我們在實驗中使用了 105bp 長的短序列(多個數據集合分別包含不同級別的錯誤)，並將這些序列比對到一個人類全基因的參考序列上。平均來講，CS4splice 能夠成功比對 427,786 個序列中的 88.44%(召回率)，但其他的軟件工具的召回率就低得多：SpliceMap 有 48.72%，MapSplice 有 58.41%，ABMapper 有 51.39%。就準確率來說，在所有的實驗中，CS4splice 的準確率也是最高的或者非常接近於最高的。由於我們設計的種子策略採用了 4 個互補性種子，CS4splice 要耗用更多的資源，是其他的工具兩倍多。實驗結果說明，CS4splice 所耗用的資源是我們可負擔的，也是值得的。