Indexing DNA Sequences Using q-Grams

Xia Cao, Shuai Cheng Li, and Anthony K.H. Tung

Department of Computer Science, National University of Singapore
{caoxia, lisc, atung}@comp.nus.edu.sg

Abstract. We have observed in recent years a growing interest in similarity search on large collections of biological sequences. Contributing to the interest, this paper presents a method for indexing the DNA sequences efficiently based on q-grams to facilitate similarity search in a DNA database and sidestep the need for linear scan of the entire database. Two level index – hash table and c-trees – are proposed based on the q-grams of DNA sequences. The proposed data structures allow the quick detection of sequences within a certain distance to the query sequence. Experimental results show that our method is efficient in detecting similarity regions in a DNA sequence database with high sensitivity.

1 Introduction

Similarity search on DNA database is an important function in genomic research. It is useful for making new discoveries about a DNA sequence, including the location of functional sites and novel repeatitive structures. It is also useful for the comparative analysis of different DNA sequences. Approximate sequence matching is preferred to exact matching in genomic databases due to evolutionary mutations in the genomic sequences and the presence of noise data in a real sequence database. Many approaches have been developed for approximate sequence matching. The most fundamental one is the Smith-Waterman alignment algorithm [14] which is a dynamic programming approach that seeks the optimal alignment between a query and the target sequence in $O(mn)$ time, m and n being the length of the two sequences.

However, these methods are not practical for long sequences in the megabases range. Effort to improve the efficiency falls into the common idea of filtering by discarding the regions with low sequence similarity. A well known approach is to scan the biological sequences and find short "seed" matches which are subsequently extended into longer alignments. This method is used in program like FASTA [13] and BLAST [1] which are the most popular tools used by biologists. An alternative approach is to build index on the data sequences and conduct the search on the index. Various index structure models [2, 4, 7, 17] have been proposed for this purpose.

Our method is based on the observation that two sequences share a certain number of q-grams if the edit distance between them is within a certain threshold. Moreover, since there are only four letters in the DNA alphabet, we know that the number of all combinations of q-grams in a DNA sequence is 4^q.

In this paper, we propose two level index to prune data sequences that are far away from the query sequence. The disjoint segments with the length ω are generated from

L. Zhou, B.C. Ooi, and X. Meng (Eds.): DASFAA 2005, LNCS 3453, pp. 4–16, 2005.

the sequence. In the first level, the clusters (called qClusters) of similar q-grams in DNA sequence are generated; then a typical hash table is built in the segments with respect to the qClusters. In the second level index, the segments are transformed into the c-signatures based on their q-grams; then a new index called the *c-signature trees* (c-trees) is proposed to organize the c-signatures of all segments of a DNA sequence for search efficiency.

In the first level of search, the sliding segment of query sequence is generated and encoded into the key in terms of the coding function, and then the neighbors of this key will be enumerated. Thus a set of candidate segments will be extracted from the buckets pointed by the key and its neighbors, and be put into the second index structure c-trees for future filtering. In the second level of search, we only access the tree paths in c-trees that include possible similar data sequences in their leaf nodes. We also propose a similarity search algorithm based on the c-trees for query segments.

The rest of paper is organized as follows. In Section 2, we define the problem of similarity search in DNA sequence databases and briefly review related work. In Section 3, the concept of qClusters and c-signature is presented. The filter principle based on q-grams is also described. In Section 4, we propose two-level index scheme constructed on the q-grams for DNA sequences. In Section 5, an efficient similarity search algorithm is presented based on the proposed index structure. The test data and experimental results are presented in Section 6. Section 7 summarizes the contribution of this paper.

2 Problem Definition and Related Work

In this section, we formalize the similarity search problem in a DNA sequence database and describe the related existing work.

2.1 Problem Definition

The problems of approximate matching and alignment are the core issues in sequence similarity search. To process approximate matching, one common and simple approximation metric is called $edit\ distance$.

Definition 1. Edit Distance
The edit distance between two sequences is defined as the minimum number of edit operations (i.e., insertions, deletions and substitutions) of single characters needed to transform the first string into the second. $ed(S, P)$ *is used to denote the edit distance between sequence* S *and* P.

In general, this problem of sequence search can be described formally as follow:

Problem 1. Given the length l and edit distance ϑ, find all subsequences S in $\mathcal{D}$ which have length $|S| \geq l$ and $ed(S, Q') \leq \vartheta$ for subsequence Q' in query sequence Q.

Since with high possibility there exists a similar segment pair $(s, q), s \in S, q \in Q'$ if S is similar to Q', we instead solve the following problem.

Problem 2. Given the length ω and edit distance ε, find all the segments s_i with length ω in $\mathcal{D}$ which meet $ed(s_i, q_j) \leq \varepsilon$ for the query segments q_j with length ω in Q.