Using Binary Decision Diagrams (BDDs) for Memory Optimization in Basic Local Alignment Search Tool (BLAST)

Demian Oliveira, Fernando Braz, Bruno Ferreira,
Alessandra Faria-Campos, and Sérgio Campos

Department of Computer Science
Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627, Pampulha, 30123-970, Belo Horizonte, Brazil
demianbueno@yahoo.com.br,
{fbraz,bruno.ferreira,alessa,scampos}@dcc.ufmg.br

Abstract. Sequence alignment is the procedure of comparing two or more DNA or protein sequences in order to find similarities between them. One of the tools used for this purpose is the Basic Local Alignment Search Tool (BLAST). BLAST however, presents limits on the size of sequences that can be analyzed requiring the use of a lot of memory and time for long sequences. Therefore, improvements can be made to overcome these limitations. In this work we propose the use of the data structure Binary Decision Diagram (BDD) to represent alignments obtained through BLAST, which offers a compressed and efficient representation of the aligned sequences. We have developed a BDD-based version of BLAST, which omits any redundant information shared by the aligned sequences. We have observed a considerable improvement on memory usage, saving up to 63,95% memory, with a negligible performance degradation of only 3,10%. This approach could improve alignment methods, obtaining compact and efficient representations, which could allow the alignment of longer sequences, such as genome-wide human sequences, to be used in population and migration studies.

Keywords: Binary Decision Diagrams (BDD), Basic Local Alignment Search Tool (BLAST), Multiple Sequence Alignment.

1 Introduction

One of the fundamental problems in Bioinformatics is the alignment of sequences, which is used to compare and find similarities between primary biological sequences, such as DNA or proteins. Several algorithms have been proposed to address this issue. One of the most used is the Basic Local Alignment Search Tool (BLAST) [1]. However, it presents some limitations regarding the size of the sequences that can be analyzed. Therefore, improvements can be made to overcome these limitations.

S. Campos (Ed.): BSB 2014, LNBI 8826, pp. 65–72, 2014.

In this paper, we present a different and novel approach to solve this problem, using the data structure Binary Decision Diagram (BDD), a special type of Binary Decision Tree (BDT), which allows a compressed, concise and efficient data representation. During the execution of BLAST, we represent the aligned sequences as a BDD, which discards redundant data shared by the sequences, which saves memory used by the program.

We have also performed a comparative study, measuring the execution time and memory usage of BLAST, available at the National Center for Bioinformatics Information (NCBI)[1], and our BDD-based version. We have observed a considerable improvement in memory usage, saving up to 63,95% memory, with a small trade-off of negligible performance degradation of only 3,10%.

This approach could improve alignment methods, obtaining compact and efficient representations, which could allow the alignment of longer sequences, such as genome-wide human sequences, to be used for population and migration studies.

2 Background

2.1 Basic Local Alignment Search Tool (BLAST)

Sequence alignment is the procedure of comparing two or more DNA or protein sequences by searching for series of individual characters that are in the same order in the sequence for the purpose of identifying similar regions, which may share characteristics, such as structure and function. Several algorithms exist to accomplish this, being prominent among them the **Basic Local Alignment Search Tool (BLAST)**. BLAST is used for the analysis, study and comparison of primary biological sequences, such as nucleotides in DNA and RNA sequences and amino acids in proteins [1,7]. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify sequences that resemble it above a certain threshold. Different types of BLAST are available according to the sequences to be compared. A BLAST search allows the user to find alignments of a source biological sequence, called **query**, with another sequence, called **subject**, aiming to infer biological homology.

Since its creation, BLAST has been extended in different ways. Parallel implementations have been created, such as PLAST [8]. Specific optimizations of its algorithms, such as improving the order of its index seed, have also been performed [6]. BLAST+ , a complete reimplementation in C++ of BLAST (originally implemented in C), has also been created [3]. There are also non-equivalent alternatives for database search, such as BLAT [5], which uses the index of all nonoverlapping K-mers in the genome.

Finally, there are some works exploring the use of GPU and CUDA cores, for intensive parallel computations, such as GPU-BLAST [9] and G-BLASTN [10]. However, the use of a data structure known as Binary Decision Diagram (BDD) to improve the algorithm efficiency has yet not been pursued.

[1] NCBI website, `http://www.ncbi.nlm.nih.gov/`