Information Sciences 297 (2015) 118–139

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

INFORMATION SCIENCES

A cascaded pairwise biomolecular sequence alignment technique using evolutionary algorithm

Gautam Garai [a], Biswanath Chowdhury [b,*]

[a] Computational Science Division, Saha Institute of Nuclear Physics, 1/AF Bidhannagar, Kolkata 700064, India
[b] Department of Bioinformatics, DOEACC Society, Jadavpur University Campus, Kolkata 700032, India

ARTICLE INFO

Article history:
Received 24 February 2014
Received in revised form 3 November 2014
Accepted 7 November 2014
Available online 15 November 2014

Keywords:
Sequence alignment
Decomposition
Cascading
Genetic algorithm
DNA/protein sequence
Experimental comparison

ABSTRACT

In computational biology, biological sequence alignment is an important and challenging task for sequence analysis. In this paper, we propose a new sequence alignment technique based on a genetic algorithm (GA) for determining the optimal alignment score for a pair of sequences that could be either DNA or protein sequences. The search space requirement of the proposed genetic-based method, named Cascaded Pairwise Alignment with Genetic Algorithm (CPAGA), is reduced by breaking a large space into smaller subspaces. This is performed by decomposing the sequence pair into multiple segments before starting the alignment procedure. Such decomposition enhances the ability of the search process to reach the global or a near-global optimal solution even for the longer sequences. The method was tested using several DNA/protein sequence pairs. We also compared the alignment score of the CPAGA with that of some well-known and relevant alignment techniques. The performance of the CPAGA method and other relevant techniques was assessed by a set of non-parametric statistical approaches, which suggest a superior performance of CPAGA over the other alignment procedures.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

Biological sequences, such as nucleotides and amino acids, accumulate mutations in the course of evolution. Some specific residues in an amino acid sequence are conserved by natural selection because they play important functional and structural roles. Knowledge of the evolutionary history and structural properties of a sequence can be useful to find related sequences. The comparison between new and existing sequences is one of the primary objectives of bioinformatics to draw inferences on evolutionary, functional, and/or structural relationships. If two sequences from different sources have more than 30% sequence identity, they are considered to be homologous, i.e., to share a common ancestral gene [39].

Sequence alignment is an essential tool in bioinformatics, in which two or more sequences are compared to align their residues (e.g., nucleotide bases of DNA and RNA, or amino acids of a protein). The optimal alignment procedure arranges sequences such that a maximum number of either identical or similar residues are matched [24]. Thus, pairwise alignment arranges two sequences in a way that maximizes their similarities or identities. Multiple Sequence Alignment (MSA) is an extension of the pairwise alignment procedure, in which more than two sequences are considered [39]. The aligned residues arranged in parallel rows can be a match, a mismatch, or a gap. Gaps are often inserted to signify insertion or deletion (indel)

---

* Corresponding author. Tel.: +91 80 1319 6284; fax: +91 33 2337 4637.
E-mail addresses: gautam.garai@saha.ac.in (G. Garai), bchowdhury2410@gmail.com (B. Chowdhury).

http://dx.doi.org/10.1016/j.ins.2014.11.009
0020-0255/© 2014 Elsevier Inc. All rights reserved.

events in a sequence alignment, and are hence assigned penalties in the scoring process. For DNA sequences, alignment scoring is used as a simple identification scheme where identical bases in both sequences are assigned positive scores. In contrast, for protein sequences, a similarity score could also be computed (in addition to sequence identity) denoting the alignment of amino acids that show similar physicochemical properties (e.g., Ser and Thr). Substitution matrices are then consulted for similarity measurement. The substitution matrices frequently used for protein sequence alignment are Point Accepted Mutation (PAM) [3] and BLOcked SUbstitution Matrix (BLOSUM) [15].

There are two main types of sequence alignment: global and local. Global alignment considers the total length of each sequence, whereas local alignment tries to align the locally highest scoring region of densely similar characters regardless of the remaining sequence length.

For a pairwise alignment problem, Dynamic Programming (DP) constructs a two-dimensional matrix where two axes represent the two sequences. DP attempts to exhaustively align all possible pairs of residues according to a scoring scheme for matches, mismatches, and gaps. The highest score is ultimately obtained by determining the optimal diagonal path by back-tracing [6,40]. The DP-based Needleman–Wunsch algorithm [26] is used for global alignment, whereas the Smith–Waterman algorithm [35] is used for local alignment. The DP has been demonstrated to produce optimal alignment [33]. The optimized alignment function is also a biologically optimal alignment, but is rarely possible when more than three sequences are considered [27]. The computation itself is also a complex task and demands computer resources when applied to MSA [37].

Therefore, to reduce this computational complexity, various heuristic approaches have been developed that can provide a solution to a problem; however, they do not guarantee finding the global optimum. BLAST [1] and FASTA [31] are the most commonly used heuristic algorithms for pairwise alignment. For MSA, three types of heuristic algorithms are generally used: progressive, iterative, and consistency-based algorithms [22,27]. Clustal W [36] is a well-known progressive alignment algorithm that is widely used for MSA. It starts with determining every possible global pairwise alignment [26] of the input sequences and then produces a distance matrix. Finally, it generates a consensus alignment by gradually adding sequences following a guide tree based on the distance matrix. However, the main drawback of this method is its so-called "greediness," in that once a sequence is aligned it cannot be altered. On the other hand, in an iterative alignment technique, the optimal solution is achieved by iteratively modifying the suboptimal solutions in the intermediate stages, which helps to solve the "greedy" nature of the progressive approaches. MUSCLE [7] is an iterative alignment method that solves the alignment problem using a profile function called the log expectation score. Another heuristic method is progressive alignment with a consistency-based scoring scheme. This scoring scheme depends on the collection of methods that simultaneously align two sequences. In the T-Coffee package [29], the collection of pairwise alignments is the combination of global (produced with Clustal W) and local (produced with Lalign [19]) alignments. For aligning every pair of residues in a sequence pair, a consistency score is estimated from the collection of methods. Since an optimal initial alignment is progressively chosen from many alternative alignments during alignment construction, T-Coffee overcomes the greedy nature of the progressive approach.

In this paper, we propose a genetic algorithm (GA)-based approach in a cascaded manner that can be used to solve the pairwise sequence alignment problem. The technique is named Cascaded Pairwise Alignment with Genetic Algorithm (CPA-GA). The novelty of this algorithm for sequence alignment is that it can be applied as an optimization tool in a large and complex search space. Its cascaded nature first breaks the large search space into several smaller subspaces by decomposing the sequence pair into multiple segments. Then it starts searching for solutions over the subspaces. As a result, the proposed technique shows good ability to move out of the local optima. The final alignment score is the summation of the best scores found in the subspaces.

The rest of the paper is organized as follows. Section 2 has the related works. Section 3 describes the conventional genetic method and its basic steps algorithmically. Section 4 elaborates the proposed technique with examples. The experimental results are reported in Section 5. The results are compared analytically and statistically with some well-known and relevant algorithms. Finally, Section 6 concludes the paper.

2. Related works

To solve sequence alignment problem in an optimized way other than the conventional approaches, several researchers have applied various evolutionary techniques which are heuristic in nature. Cutello et al. [2] developed a hybrid bio-inspired algorithm known as Immunological MSA algorithm (IMSA) which behaved as an improver to refine the best initial alignment produced by Clustal W. Wang and Li [38] proposed an iterative optimized algorithm to move blocks of gaps in MSA to refine the aligned positions. Othman et al. [30] developed a GA-based technique for pairwise nucleotide sequence alignment where each alignment was represented as a two-row matrix, but the method could not ensure the optimal solution. Jangam and Chakraborti [21] proposed a pairwise nucleic acid alignment method using the Ant Colony Optimization (ACO) and the GA. The sets of sequences were aligned by the ACO technique and then further processed by the GA. It performed better for smaller sequences, but deviated in performance for larger length sequences. The popular software developed by Notredame and Higgins is SAGA [28]. This was developed for MSA with a GA having 22 different types of complex operators used to obtain optimum alignment by using two different objective functions. The primary drawback of this algorithm was that it was time consuming due to the repeated use of the fitness function. Gupta et al. [14] also applied a GA for MSA using DNA families of *Canis familiaris*. Shyu and Foster [34] developed a DNA sequence alignment technique that optimized the

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات