Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

An efficient algorithm to detect common ancestor genes for non-overlapping inversion and applications

Fatema Tuz Zohora *, M. Sohel Rahman

AℓEDA Group, Department of CSE, Bangladesh University of Engineering and Technology, Dhaka-1000, Bangladesh

ARTICLE INFO

Article history:
Received 9 July 2015
Received in revised form 24 February 2016
Accepted 5 March 2016
Available online xxxx

Keywords:
Consensus string
Inversion
Genetic disease
Allelic heterogeneity
Iterative algorithm
Matrix

ABSTRACT

In this paper, an algorithm is proposed that detects the existence of common ancestor gene sequences for non-overlapping inversion (reversed complement) metric given two input DNA sequences. Theoretical worst case running time complexity of the algorithm is proven to be $O(n^4)$, where n is the length of each input sequence. However, by experiment, the running time complexity is found to be $O(n^3)$ for the worst case and $O(n^2)$ for average case. Moreover, the worst case occurs when both input sequences have the similarity of around 90% which is very rare. This work is motivated by the purpose of diagnosing an unknown genetic disease that shows *allelic heterogeneity*, a case where a normal gene mutates in different orders resulting in two different gene sequences causing two different genetic diseases. Our algorithm can potentially save huge energy and cost of the existing diagnostic approaches. The algorithm can be useful as well in the study of breed-related hereditary conditions to determine the genetic spread of a defective gene in the population.

1. Introduction

Computer Alignment of molecular sequences is widely used for biological sequence comparisons. Genetic diseases are caused by gene mutation, which can be inherited through generations and can result in new sequences from a normal gene [16]. In this paper, we propose a deterministic algorithm for detecting the existence of ancestor gene sequence(s) directly from two given mutated gene sequences. In particular, we focus on a genetic disorder called *allelic heterogeneity* as a domain of application, and to the best of our knowledge this is the first attempt to do so. Allelic heterogeneity is considered to be the greatest challenge for molecular genetic diagnosis as stated in the book by Meisenberg et al. [11]. It is a case where different mutations in the same gene result in different phenotypes which may lead to diseases with entirely different clinical features [12]. For example, mutations in the *IDUA* gene have been implicated in the etiology of *Hurler syndrome* disease as well as of *Scheie syndrome I*. So these two are called allelic heterogeneous to each other. Allelic heterogeneity motivates us with its importance in medical science. It also causes autism and rigid-compulsive behaviors [14]. Very recently Castellani [5] presented CFTR2, a novel approach for the clinical diagnosis of genetic disorders emphasizing specially the allelic heterogeneity. Detection of an unknown disease as allelic heterogeneous with a known genetic disease helps in medication and treatment. For this purpose whole genome sequencing is required which takes around 12 to 13 weeks (data collected from https://www.genetests.org). Besides, diagnosis of such disease needs approaches like mismatch scanning, gene sequencing, linkage analysis etc., all of which are highly expensive solutions as apparent from the cost

* Corresponding author.
E-mail addresses: anne.06.cse@gmail.com (F.T. Zohora), sohel.kcl@gmail.com (M.S. Rahman).

http://dx.doi.org/10.1016/j.tcs.2016.03.006
0304-3975/© 2016 Elsevier B.V. All rights reserved.

Table 1
Gene name with corresponding allelic heterogeneous diseases and diagnosis details (data is collected from: https://www.genetests.org/tests; http://www.ggc.org/) last accessed: 28–01–2016.

Gene name	Allelic heterogeneous disease		Diagnosis details		
	Disease 1	Disease 2	Diagnostic method	Cost	Time
IDUA	Hurler syndrome	Scheie syndrome I	Sequencing	$2050	3 weeks
CFTR	Cystic fibrosis	Congenital absence of the vas deferens	Sequencing	$1310.00	3 to 4 weeks
DMD	Duchenne muscular dystrophy	Becker muscular dystrophy	MLPA	$500	2 weeks
RET	Hirschsprung disease	Multiple endocrine neoplasia Type 2	Sequencing	$1160.00	3 to 4 weeks

estimates provided in Table 1. For example, diagnosis of Hurler syndrome or Scheie syndrome I takes three to four weeks with gene sequencing approach and costs around $2050. Since the clinical diagnosis is extremely expensive it is worth investigating whether a tractable/polynomial time algorithm exists to detect the possibility of allelic heterogeneity. And this is the main goal of this paper. Here we use the term *common ancestor* to indicate the same gene sequence from which different mutation order gives different gene sequences x and y. Detecting allelic heterogeneity can be formally defined as follows:

> Given x and y as input, where x is the gene sequence of a known disease caused by mutation of some ancestor gene p, and y is the gene sequence of an unknown disease, if there exist p as a common ancestor between x and y, then we can diagnose that unknown disease y to be allelic heterogeneous to x.

In this paper we work on inversion mutation and intend to work on other mutations in future.

The Consensus problem in strings is motivated by the requirement of finding commonality of a large number of strings and has a variety of applications in bioinformatics [7]. It has biological applications concerning signal finding and primer design [10]. Amir et al. [1] proposed a new pattern matching paradigm *Pattern Matching with Rearrangements* being motivated by the *Sorting by Reversals* problem [3,4]. The consensus string problem is defined as follows:

> Given a set of k strings $S = \{s_1, \ldots, s_k\}$ and a constant d, find, if it exists, a string $s^\star$ such that the distance of $s^\star$ from each of the strings in S does not exceed d, for some suitable and meaningful definition of the term 'distance'.

We can map the problem of determining allelic heterogeneity to a *relaxed* version of the consensus string problem as follows. In allelic heterogeneity, a perfect gene p is mutated in different order and gives different sequences x and y. On the other hand, in consensus string problem, we have x and y as input, and we have to go on reverse direction by applying the mutations over x and y to reach p, a candidate consensus string. Here p need not be away by the minimum distance d from x and y. That is, we need to drop the constraint of the minimum distance d and just find out whether p exists as a candidate consensus string (common ancestor). Since the minimum distance d is not present as a parameter, our problem can be thought of as a *relaxed version* of the original consensus string problem. Our proposed algorithms can be summarized as follows:

- We first propose an algorithm for determining the *existence* of one or more candidate consensus strings ($s^\star$), given two strings x and y of length n on an alphabet of size $k = 4$ under the distance metric called *non-overlapping inversion*, i.e., reversed complements.
- Later we show how this algorithm can be used to determine whether there exists a specific gene sequence p as the common ancestor of input gene sequence x and y, i.e., whether x and y are allelic heterogeneous to each other.
- At the end we also show the approach for producing all possible common ancestors to make the algorithm applicable in 'breed-related hereditary conditions'.

The problem of detecting the existence of common ancestors has been introduced very recently by Cho et al. [8]. In particular, they have provided an $O(n^3)$ algorithm using $O(n^2)$ space, where n is the size of the two input strings. But unfortunately we have found through experimentation that their algorithm fails in returning the correct answers in some cases because of not tracking the prefixes of the common ancestors. We will explain the error along with a simulation of a counter example in a subsequent section. In this paper, we present a new algorithm which correctly solves the problem with the same time complexity in the worst case. We further present experimental evidence that our algorithm in practice runs in quadratic time. Although the motivation behinds the work of Cho et al. [8] was to examine the problem of deciding whether or not two gene sequences are mutated from the same gene sequence, but they could not identify specific domain of application where that finding can be used. But we target allelic heterogeneity, a biomedical problem. This definitely makes the problem more interesting and particularly extremely useful.

The consensus string problem under the *reversal metric* is in general NP hard according to the work of Amir et al. [2]. However, Gramm et al. [9] have shown that exact solutions for consensus string and related problems exist for constant distance parameter (d) and constant number of strings (k). Since alignment with inversions in general does not have a known polynomial time algorithm, a simplification to the problem considers only non-overlapping inversions (each index