Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Genomics

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

Parallel approach on sorting of genes in search of optimal solution

Pranav Kumar*, G. Sahoo

Birla Institute of Technology, Mesra, India

ARTICLE INFO

Keywords:
Sorting by transposition
Sorting by fission
Algorithm
Sorting by fusion
Genome rearrangement

ABSTRACT

An important tool for comparing genome analysis is the rearrangement event that can transform one given genome into other. For finding minimum sequence of fission and fusion, we have proposed here an algorithm and have shown a transformation example for converting the source genome into the target genome. The proposed algorithm comprises of circular sequence i.e. "cycle graph" in place of mapping. The main concept of algorithm is based on optimal result of permutation. These sorting processes are performed in constant running time by showing permutation in the form of cycle. In biological instances it has been observed that transposition occurs half of the frequency as that of reversal. In this paper we are not dealing with reversal instead commencing with the rearrangement of fission, fusion as well as transposition.

1. Introduction

Gene matching is a process through which a portion of chromosome is feed into the algorithm as an input in order to find out the desired output based on the best possible match among the several gene sequences available in the database. Chromosomes are the combination of genes, comprises of packed thread like structure of DNA molecules present in every cell of living organisms and genome [11] is a set of chromosomes. Cellular DNA provides instructions for building various proteins which is necessary for the cell to survive. DNA is a double stranded molecule in which each strand is a long sequence of four types of nucleotides – adenine (A), thymine (T), cytosine (C) and guanine (G). When two long chains of nucleotides twisted in the form of double helix structure joined by hydrogen bond is called DNA strand. The two strands are called complementary strands, as one strand can be defined from the other strand. Ordered set of genes is called genome. Genome rearrangement is defined as a segment of gene flips and reinserted in specific location of gene sequence in a same or different chromosome. Here we are dealing with three basic genome rearrangement process i-e fission, fusion and transposition. It can be noted that fission and fusion [17] comes under multi chromosome operation and transposition is a single chromosome operation. In transposition a gene sequence from a chromosome breaks at three points and the segment between any two positions is removed and reinserted onto the third position. Fission is a multi-chromosome operation in which a single chromosome splits up into two or more chromosomes, whereas in fusion two or more chromosome merges together to form a single chromosome. Multi chromosome operation means the rearrangement process occurs between different chromosomes. Mira and Meidanis [17] have used the concept of reversal along with fission and fusion whereas here we are dealing with transposition instead of reversal [2,10] not only to make the algorithm faster, it intern leads to inherent complexity by performing concurrent tasks. As transposition sort is originally developed for the use of parallel processors with local interconnections and here also our work is based on parallel selection of transposition points.

Apart from this it is relatively a simple algorithm for sorting. It is shifting pair wise the wrong order elements and also its better complexity is still open. Therefore we are dealing transposition instead of reversal.

A genome rearrangement is a genomic mutation as a result of which it is converted into new species such as cabbage can be converted to turnip. These large scale changes can be harmful but sometime it provides significant advantage like species recognition. Scientists believe that instead of studying the whole gene sequence, it can be done by the rearrangement of a portion of genome. Which will help them understand how genome as a whole works, how gene works together for the development, growth and maintenance of organism. Also previously there was a lack of portable algorithms and what so ever exists are more compatible with the uni-chromosomes and not fully with the multi chromosomes, so our algorithm adds on to it, which results in the existence of parallelism and also help other people to understand that to work with transposition is easy. In this paper, we revisit the problem of sorting by transposition and fission [14,15] and fusion along with that finding an optimal solution, for that we present a new algorithm that traverse traces [13] using three sorting operations having two or more processes working in parallel, would be faster enough than the

* Corresponding author.
E-mail address: pranav5503.06@bitmesra.ac.in (P. Kumar).

http://dx.doi.org/10.1016/j.ygeno.2017.09.006
Received 26 May 2017; Received in revised form 18 August 2017; Accepted 12 September 2017

sequential operation and also more powerful.

2. Notations and nomenclature

In this section we have discussed the analysis of sorting by fission, fusion, transposition and some concept of reversal and transposition also. A set of gene form a chromosomes and they combined together to form genome. Here we mainly deal with signed permutation [8,9] that means the sign of genes follow some specified directions shown by positive or negative direction. If (a, b, c, d), (e, f) and (g, h) are written in three rows this means each row denotes a chromosomes and all together form a genome in other words there are eight genes a, b, …h.

Two genomes are said to be similar if the set of chromosomes are similar. As genes are the combination of nucleotides A, C, T, G in a repeated pattern, so it will be difficult to judge which combination is moved to new position. To resolve this issue we have represented genes with signed integer [12]. While plotting a cycle graph [3], we assign a direction for specifying the forward or backward movement, shown by an arrow. Let us assume a genome.

A = ((a^{11}, a^{12}, …,a^{1n}), (a^{21}, a^{22}, …,a^{2n}), … (a^{k1}, a^{k2}, …a^{kn})) and B = ((b^{11}, b^{12}, …,b^{1n}), (b^{21}, b^{22}, …,b^{2n}), … (b^{k1}, b^{k2}, …b^{kn})), where mathematical representation of 'A' specifies starting genome and 'B' represent goal genome, but with a condition that both starting and goal will consist of same group of genes and no gene is in repetition form. Signed permutation is shown by positive or negative sign, for arbitrary sequence M = m^1, m^2, …, m^k is represented in reverse as M = $-m^1$, $-m^2$, …,-m^k. A chromosome Z is indistinguishable if either Z = X or Z = −X same as genome A and B indistinguishable if genes of chromosomes of A and B are same. As in case of {(1, 2, −3), (4, −5, 6, 7), (8, 9)} represent 9 genes from 1 to 9 in three rows called chromosome. Here we compare the elements of each chromosome of first genome with the goal genome and if a position variation is found within or outside the chromosome, then we reverse the gene within the chromosome or swap the elements with other chromosome. In transposition [6] a portion of chromosome changes their position. For example, if (ρ^{i1}, ρ^{i2}, ρ^{i3}, ρ^{i4}, …, ρ^{ik-1}, ρ^{ik}) be the sequence of gene then it can be written as (ρ^{k-1}, ρ^{i4}, ρ^{i1}, ρ^{i3}, …, ρ^{i2}, ρ^{ik}). In this paper we mainly focused on finding optimal solution among several, so for optimal solution [1] a given source genome to be converted into target genome with a minimum number of steps that can be measured by finding permutation distance [5] and can be achieved by forming a graph.

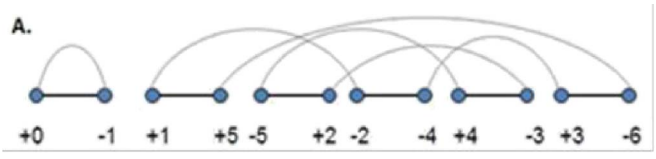
2.1. Chart

A cycle graph is a cycle of alternate grey and black edge path. Hannenhali and Pevzner [4] showed that every reversal increases the parameter c (number of disjoint cycles) at most by one, so number of reversals required to sort a permutation of length n is given by d ≥ n + 1 − c since breakpoint graph of identity permutation of length n has (n + 1) disjoint cycles.

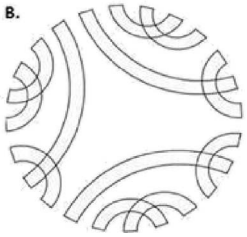
Fig. 1(a) is the breakpoint graph [4] and Fig. 1(b) overlap graph of π is constructed by joining two connectivity between pair of elements, as the neighboring adjacent elements are connected by a black edge and a grey edge from i to (i + 1) shown with a dotted line connectivity. This convention changed in case of signed permutation, in which a grey edge is connected between +i to −(i + 1). It is a convention of dotted arcs extending above those black edges. The two color (black, grey) cycle graph of any permutation say A and B (where A is the source genome and B is the target genome) be denoted as G (A, B) where set of vertices is from 1 to n in some particular order (i^-, i^+ or i^+, i^-) here n is the number of genes.

3. Obstacles and bastions

In graph shown in Fig. 2 when two alternate black edges is connected with the same grey edge, overlapping two cycles is called as



(a) shows breakpoint graph



(b) overlap graph

Fig. 1. (a) shows breakpoint graph. (b) overlap graph.

Fig. 2. A is obstacle, B is non-obstacle and C is super obstacle.

obstacle and when a obstacle is within another obstacle known as super obstacle. If any other cycle exists between a obstacle said to be a non-obstacle. Obstacle is a bad component in a graph.

If there are exists an odd number of super obstacles in a cycle it is called as bastion. Each and every oriented component can be classified as either a obstacle or a protected non-obstacle given by Siepel [16]. A protected obstacle is the one which separate other unoriented components; whereas simple obstacle does not. If a traversal is possible to pass in either direction from second vertex to the third vertex without encountering a first vertex, while separation used with unoriented component the definition applies as well to benign components. Let grey edge (u, v) of graph G(A, B) be considered as internal if u, v belongs to the same chromosome i.e. A; if different is external. If two genes are connected with different signs in A then the internal grey edge is called to be oriented; unoriented if they are not oriented, except when they are trivial. A trivial cycle consists of a single black edge, grey edge and isolated vertex in Fig. 1(b). Let N = {n_i} be the set of components, C_i = {$c_{i,j}$} the set of cycles corresponds to n_i and $B_{i,j}$ = {$b_{i,j,k}$} the set of black edges corresponds to $c_{i,j}$. In breakpoint graph [6] flow of connectivity is shown by convergent or divergent direction, so $b_{i,j,k}$ and $b_{i,j,1}$ must be convergent or divergent if any two black edges correspond to the same cycle. If divergence exists in black edges $b_{i,j,k}$ and $b_{i,j,1}$ then set of cycle $c_{i,j}$ is oriented; otherwise unoriented until and unless $B_{i,j}$ = 1, in case $c_{i,j}$ is trivial. If $c_{i,j}$ is oriented and $c_{i,j} \in C_i$, then m_i is oriented; otherwise unoriented except $|C_i|$ = 1 and if it is trivial then m_i is also trivial [16].

4. Fission, fusion and transposition process

A fission for a sequence π = (π^1… π^i… π^{i-1}… π^j… π^{j-1}… π^k… π^{k-1}… π^n) splits up into three different sequences i-e (π^1… π^i… π^{i-1}), (π^j… π^j) and (π^k… π^{k-1}… π^n). Whereas in fusion different sequences i-e (π^1… π^i… π^{i-1}), (π^j… π^j) and (π^k… π^{k-1}… π^n) merges together to form a single sequence π = (π^1… π^i… π^{i-1}… π^j… π^{j-1}… π^k… π^{k-1}… π^n). A transposition μ (i, j, k) for a sequence π = (π^1… π^i… π^{i-1}… π^j… π^{j-1}… π^k… π^{k-1}… π^n), where 1 ≤ i < j < k ≤ n + 1, cuts an interval point from [i, j-1] and shifted to another position onto the same sequence as πμ =