Computers in Biology and Medicine 48 (2014) 109–118

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Subdividing globally important zones based on data distribution across multiple genome fragments

Feng Chen [a,b], Yuhong Zhang [a], Yi-Ping Phoebe Chen [b,*]

[a] College of Information Science and Engineering, Henan University of Technology, Zhengzhou, China
[b] Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, Australia

ARTICLE INFO

Article history:
Received 19 August 2013
Accepted 7 February 2014

Keywords:
Globally important zone
Common insertional sites
Insertion distribution
Multiple genome fragments
Hierarchical and density-based method

ABSTRACT

In multiple genome fragments, a globally important mode is a zone represented by a significant change, where the change has a similar impact on every related fragment in the zone. This zone may represent the cancer related genes involved in diverse tumors. Globally important zones are characterized by two features: (1) there are more data points in globally important zones than in other areas of fragments; (2) the data points are distributed evenly on as many genome fragments as possible. Globally important zone mining needs to contain the following features: (1) independent of data distribution; (2) noise filtering; (3) pattern boundary identification; and (4) zone ranking. We have developed a hierarchical and density-based method, called GIZFinder (globally important zone finder), to detect and rank such zones based on two criteria: distribution width and distribution depth. The comparisons on the simulated data shows our method performs significantly better than the kernel framework and the sliding window. By experimenting on real cancer gene data, we identify 53 novel cancer genes, some of which have been proven correct.

1. Introduction

There is a lot of research on multiple genome fragments or data streams [1–10]. For example, change detection focuses on how the data changes among streams [1]. Frequent episodes detect short ordered sequences in event streams [6]. Fast time series evaluation clusters the similar streams [9]. Different from these, globally important zones (GIZs) proposed in this paper represent a new pattern analysis technique focusing on the significant and mutual features of data points on multiple genome fragments (or data streams). Definition 1 is the formal definition of GIZs.

Definition 1. Globally important zone: let $X=\{x_1, x_2, ..., x_m\}$ be a multiple genome fragments. One fragments $x_i=\{x_{i1}, x_{i2}, ..., x_{in}\}$ $(i=1...m)$. x_{ij} represents the jth location on genome i. $\delta(j_1, j_2)=\{x_{ij} | i \in 1...m, \quad j \in j_1...j_2, \quad 1 \le j_1 < j_2 \le n\}$ is a rectangle zone across all the genome fragments. Given a measure ψ and a threshold ξ, $\delta(j_1, j_2)$ is a GIZ if $\psi(\delta(j_1, j_2)) > \xi$.

As shown in Definition 1, a GIZ indicates a rectangle zone across multiple genome fragments with the following features:

(1) The zone includes more data points than others. Statistically, a zone with more data is more significant than others as more data might imply a more important change.
(2) The number of data points on every fragment is similar. Mathematically, the distribution of data points across all the genome fragments possesses as small deviation as possible.

GIZs can be applied widely. For example in cancer gene detection, retroviral insertion is one of the main causes of gene mutation. Retroviruses insert their own DNA into the host cell's genome. Cancer could occur if the positions where the retrovirus inserts are related to cancer genes. The insertional zones that are related to cancer are called CIS (common insertional site) [12,13]. An important CIS has two features: more insertions than other zones and even distribution on as many independent genome fragments as possible. So, cancer gene detection can be transferred to find the important CISs on genome fragments [12–15]. Fig. 1 shows a real example verifying the relation between CISs and GIZs.

In Fig. 1, the eight gray lines indicate eight genome fragments from different tumor types. The hexagons represent the insertions. The zones indicate three genes related to cancer demonstrated by RTCGD (Retrovirus Tagged Cancer Gene Database). According to insertion quantity, *Fgf3* is the most important as it has 38 insertions. *Ccnd1* and *Notch1* are equally important because both of them have 33 insertions. However, Fig. 1 explains the disadvantage of simply

* Corresponding author.
E-mail address: phoebe.chen@latrobe.edu.au (Y.-P.P. Chen).

http://dx.doi.org/10.1016/j.compbiomed.2014.02.004
0010-4825

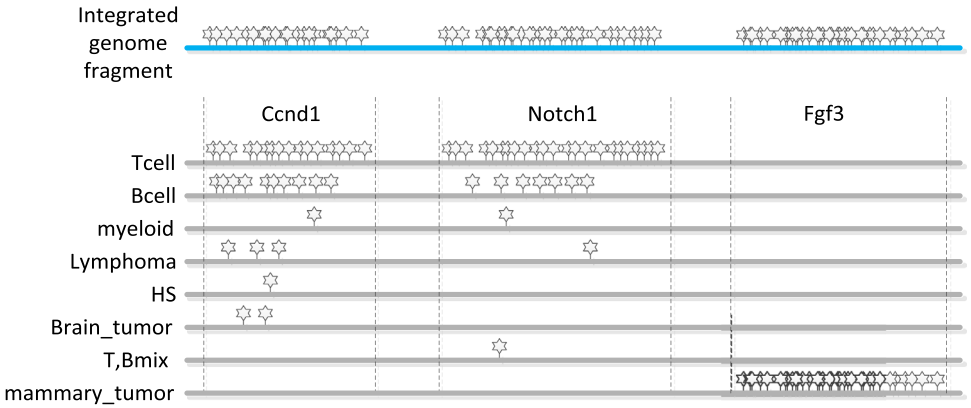

Fig. 1. An example of globally important zone in cancer gene discovery.

taking insertion quantity into consideration. All the insertions in *Fgf*3 are located on a mammary_tumor, implying that it either is unique for a mammary_tumor, or it might be a statistic error. In addition, we think *Ccnd*1 is more important than *Notch*1 because not only does *Ccnd*1 involve more tumor types than *Notch*1, but also the insertions in *Ccnd*1 are distributed more evenly than those in *Notch*1. So, *Ccnd*1 plays a more important role in more tumors. Therefore, the genes can be distinguished from each other effectively due to the consideration of insertion distribution across genome fragments.

To sum up, genes can be subdivided more effectively by combining data quantity and data distribution, i.e., two features of GIZs. Features 1 and 2 are represented by DW (distribution width) and DD (distribution depth), which indicate the length and the width of a rectangle zone, respectively. Then the measure ψ in Definition 1 is actually the zone's area, calculated by $DW \times DD$.

However, many challenges remain in GIZ detection. Firstly, the detection should be independent of data distribution. For example for CIS detection, all the genome fragments are extracted from a variety of patients or animals, which impossibly conform to a known distribution. Secondly, noises inevitably occur in a huge database. For instance, in carcinogenesis, the genes close to transcription start sites are easier to be altered [15], which could become false CISs. So filtering noise is necessary to improve detection quality. Thirdly, the GIZs without clear boundaries are vital. For instance, if a CIS's boundaries are extended incorrectly, more unrelated genes will be obtained, causing the detection of incorrect genes and misleading drug discovery [16–18]. Fourthly is ranking GIZs. It is impossible to focus on every zone due to the dramatic growth of data size. So, the most effective way is to rank all the GIZs reasonably and explore the top ones.

Based on these challenges, we have developed a series of novel techniques, called GIZFinder, to discover GIZs. Firstly, a hierarchical tree based on DBScan is built to cluster data [19]. Secondly, we have designed a novel concept: ID (Insertion distribution), combining DD and DW, to evaluate every cluster in the hierarchical tree. Thirdly, we have designed two evaluations (global and individual) to demonstrate the significance of GIZs. Finally, we have introduced four novel criteria to verify whether or not GIZFinder can identify the boundaries of GIZs. GIZFinder not only performs better than a kernel framework and a sliding window on the simulated data, but also highlights the most significant genes that occur in a variety of genome fragments on real cancer gene data.

This paper is organized as follows: Section 2 reviews the related work. Section 3 details GIZFinder. The experiments on the simulated and the real databases are analyzed in Section 4. Section 5 is the conclusion.

2. Related work

We firstly introduce the related algorithms for identifying CISs because we use GIZFinder to detect CISs in our experiments [12–15,20]. Sliding window, which has been used to build the retroviral insertion database RTCGD, defines three fixed window widths to detect CISs [12]. However, fixed window widths cannot deal with variable ranges of CISs, which is important from the biological point of view. The Kernel framework improves CIS discovery dramatically [15,20]. Using changeable window widths, CISs can be detected in any biological range. However, it is still impacted by data distribution. In this paper, we will compare these two methods with GIZFinder to prove our method is more promising. In addition, when the data size is not large, Monte Carlo simulation performs much better [13]. But as the database size increases, it could lead to more and more FPs. To some extent, Poisson distribution resolves this issue [14]. However, it is impacted by data distribution and noises too because of using Poisson distribution. In addition, these methods just take the positions of the potential cancer genes into consideration because so far, gene expression during carcinogenesis is not taken into consideration here. Summarizing, the principle of these algorithms is to integrate all the genome fragments into one long genome for data quantity analysis. However, they fail to evaluate the second feature: data distribution.

In addition, GIZs are different from co-occurrence patterns [38–41]. Co-occurrence patterns indicate a phenomenon caused by the cooperation of two or more data. For example, the database in [38] published a variety of mutation that occurred in diverse tumor types. The potential co-occurrence target genes in gastric cancer were analyzed in [38–41]. Compared with co-occurrence patterns, GIZs consider not only data quantity but also data distribution. If the insertions in one zone are distributed on a variety of the fragments, then GIZFinder will consider it a potential GIZ. However, GIZFinder will still allocate a zone only related to few fragments with a high weight as long as it includes enough insertions. Summarizing, GIZFinder focuses on highlighting the most significant zones based on more reasonable ranking.

To demonstrate that GIZFinder can be applied widely, we also review some work regarding change detection, time series and other fields. Change detection, based on density, identifies if a given baseline dataset has the same distribution with a new observed data set [1,35–37]. It focuses on “recent” changes, leading to the failure of the exploration of the previous changes. In addition, this method cannot determine the change boundaries. The detection of spatial clusters assumes that the data confirms to Possion or Bernoulli distribution [3]. It can detect clusters of any shape, which is different from our goal because a GIZ indicates