Computational Biology and Chemistry 51 (2014) 83–92

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

Determining common insertion sites based on retroviral insertion distribution across tumors

Feng Chen[a,b], Zhoufang Li[a], Yi-Ping Phoebe Chen[b,*]

[a] College of Information Science and Engineering, Henan University of Technology, Zhengzhou City, Henan Province 450001, China
[b] Faculty of Science, Technology and Engineering, La Trobe University, Melbourne, Victoria 3086, Australia

ARTICLE INFO

Article history:
Received 31 December 2013
Received in revised form 24 February 2014
Accepted 3 March 2014
Available online 12 March 2014

Keywords:
Common insertion site
Retroviral insertion
DBScan
Normal standard deviation

ABSTRACT

A CIS (common insertion site) indicates a genome region that is hit more frequently by retroviral insertions than expected by chance. Such a region is strongly related to cancer gene loci, which leads to the detection of cancer genes. An algorithm for detecting CISs should satisfy the following: (1) it does not require any prior knowledge of underlying insertion distribution; (2) it can resolve the insertion biases caused by hotspots; (3) it can detect CISs of any biological width; (4) it can identify noises resulting from statistic mistakes and non-CIS insertions; and (5) it can identify the widths of CISs as accurately as possible. We develop a method to resolve these difficulties. We verify a region's significance from two perspectives: distribution width and distribution depth. The former indicates how many insertions in a region while the latter evaluates the insertion distribution across the tumors in a region. We compare our method with kernel density estimation and sliding window on the simulated data, showing that our method not only identifies cancer-related insertions effectively, but also filters noises correctly. The experiments on the real data show that taking insertion distribution into account can highlight significant CISs. We detect 53 novel CISs, some of which have been proven correct by the biological literature.

1. Introduction

Mutagenesis resulting from retroviral insertion (Oja et al., 2007) is one of main causes of carcinogenesis (Uren et al., 2008; Mikkers and Berns, 2003; Lewinski et al., 2006). By infection, retroviruses insert their own DNA into the host cell's genome, which could lead to gene mutation of the host cell. The gene, which is close to or includes the location in which the retrovirus inserts, may be altered. If this gene is an oncogene or tumor suppressor gene (Liu et al., 2009; Tran et al., 2008), such mutation leads to the proliferation of cells without control. Finally, a tumor could develop (Miething et al., 2007; Suzuki et al., 2006; Slape et al., 2007).

The process of tumor development actually involves multi-genes and stages. In this process, the tumor tissues with the retroviral insertions are copied many times while those without the retro viral insertions might be copied a few times. As a result, research on tumor tissue can find genome regions in which there are many retroviral insertions. Therefore, the genes corresponding to these regions are very probably associated with carcinogenesis. A region is called CIS (common insertion site) if it meets two criteria: (1) it includes more insertions than can be expected by chance; and (2) the insertions are distributed across multiple independent tumors. Based on the definition and features of CISs, CIS detection should resolve the following challenges:

(1) *Insertion distribution*: There is no evidence demonstrating that retroviral insertions conform to any known distribution. So, a CIS detection algorithm with a known underlying distribution may not be able to identify real CISs.
(2) *Insertion biases*: The biological experiments show that retroviral insertions are not absolutely random. Some insertions favor distinct genes or loci, called hotspots (Nielsen et al., 2005; Hematti et al., 2004; Wu et al., 2003). For example, 25% of MLV integration is detected near transcription start sites (Lewinski et al., 2006). A hotspot related to a CIS, which also leads to gene alteration, is always detected near the genes that they impact, therefore it is important that CIS detection should be not based on any underlying distribution because an assumed distribution may filter useful hotspots.
(3) *Biological variance*: A CIS could impact multiple genes that are far from or close to it. In turn, a gene could be impacted by multiple CISs. So far, the only thing which can be confirmed is that there is no a fixed CIS width available for any biological environment. So, when detecting CISs, the algorithm should provide

* Corresponding author. Tel.: +61 3 9479 6768; fax: +61 3 9479 3060.
E-mail address: phoebe.chen@latrobe.edu.au (Y.-P.P. Chen).

http://dx.doi.org/10.1016/j.compbiolchem.2014.03.001
1476-9271/© 2014 Elsevier Ltd. All rights reserved.

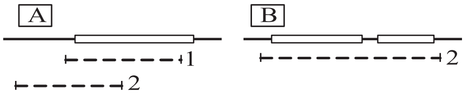

Fig. 1. Taking TP/FP as the only criterion in CIS detection. The black line represents a part of a tumor. The rectangles indicate the real CISs. The broken lines represent the CIS boundaries identified by algorithms 1 and 2. According to TP/FP, 1 and 2 perform the same because both can find the real CIS in sub-figure A. However, it is obvious that algorithm 1 is much better than 2 because the CIS from it is closer to the real CIS. In sub-figure B, 2 is considered effective because it can detect two real CISs without FPs. But we believe it is not satisfying because it combines two CISs together, which will lead to a misunderstanding of gene functions.

for candidate genes coming from any biologically relevant CIS width.

(4) *The identification of noises*: Noises in a CIS dataset can be divided into two categories: statistical errors and non-CIS insertions. If a noise, which we call noise I, is either a statistical error or a non-CIS insertion that happens in the later stage of tumorigenesis, it should be filtered. If a noise, which we call noise II, is a non-CIS insertion that happens in the early stage of tumorigenesis, it will be copied many times on the same tumor. Therefore, it can become a false CIS, which should be identified.

(5) *The detection of CIS boundaries*: Although the exact detection of CIS boundaries is still impossible, identifying CIS boundaries as much as possible is very important for understanding CIS functions and gene features. Taking TP/FP (True Positive/False Positive) as the only criterion cannot effectively identify CIS boundaries, as shown in Fig. 1. From the point of view of data mining, identifying boundaries indicates that the algorithm can match the original data exactly. So, a CIS detection algorithm should be able to match the dataset maximally. In other words, it can correctly classify as many insertions as possible.

A large volume of literature identifies CISs by assuming to integrate all the insertions on tumors into one long genome, as shown in Fig. 2A and B (Ridder et al., 2006; Akagi et al., 2004; An et al., 2005; Mikkers et al., 2002; Suzuki et al., 2002; Zhang et al., 2005; Pyrz et al., 2010). In A, retroviral insertions distribute across five tumors. For detecting CISs, all of them are integrated into one long genome to identify the regions in which there are many more insertions than others. These regions are considered to be CISs, so the genes which are close to or include them are potentially cancer related. The significance of CISs is determined by the number of insertions (Akagi et al., 2004). So CIS 2 and 3 in (B) are significant.

According to this thinking, the Sliding window (Akagi et al., 2004) uses sliding window to construct a fixed range (30k for 2 insertions, 50k for 3 insertions, 100k for 4 or more insertions). This method does not require a pre-defined parameter and is independent of insertion distribution. But it does not provide enough analysis about CIS width and insertion biases. When the size of the data is not very big, Monte Carlo simulation (Suzuki et al., 2002) and Poisson distribution (Mikkers et al., 2002) can perform very well. But these methods do not take noise and insertion biases into consideration (Nielsen et al., 2005; Hematti et al., 2004; Wu et al., 2003). The kernel framework resolves the first three challenges effectively (Ridder et al., 2006; Uren et al., 2008). It is not only independent of insertion distribution, it can also identify hotspots by smoothing the difference between insertions. Simultaneously, the window width can employ any CIS width which is biologically meaningful. So, it can also analyze CIS with distinct biological variances.

However, all these methods only focus on situation 1 being a significant CIS. They do not take insertion distribution across tumors into account. In addition, they do not consider challenges 4 and 5. First, they do not have a method to filter noise I. Second, they can not find false CISs caused by noise II because they ignore insertion distribution. Last, they take TP/FP as the only criterion, which fails to identify CIS boundaries. Based on the above analysis, we have developed a method, called DNSD (DBScan and normal standard deviation), to meet all five challenges. A stochastic review of this method is shown in Fig. 2A and C.

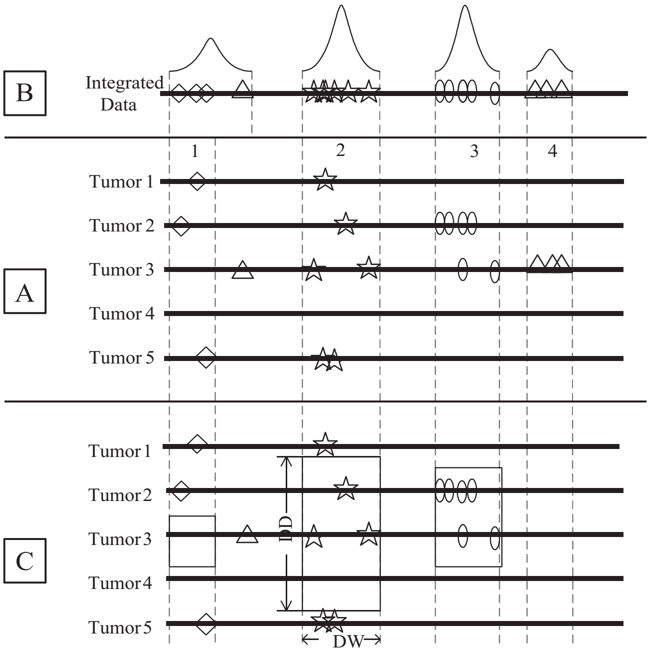

Fig. 2. Schematic view of insertion distribution across tumors, using a traditional CIS exploration algorithm and DNSD. (A) A random insertion distribution across five tumors. The diamonds, stars and ellipses are CIS-related insertions. The first triangle indicates noise I, the second three triangles in interval 4 indicate a false CIS caused by noise II. Intervals 1–3 between the broken lines indicate 3 real CISs. (B) Traditional CIS identification strategy. All the insertions are integrated into one assumed genome. The intervals between the broken lines are detected CISs. The curves indicate how many insertions in CISs. The CIS with more insertions has a higher curve. (C) DNSD introduction. Distribution Width (DW) refers to the number of insertions in CIS. Distribution depth (DD) shows how insertions distribute across tumors. The four rectangles indicate DD × DW. A CIS with a large rectangle is more important than the others. Interval 4 is filtered as a false positive.

We consider a region to be a CIS based on two criteria: distribution width (DW) and distribution depth (DD). Firstly, DW is evaluated, which indicates the number of insertions in a region. It meets criterion 1 of being a CIS, which is consistent with the current CIS detection algorithms. DD, that is for criterion 2, evaluates the distribution of the insertions across the tumors. DBScan is used for calculating DW and filtering noise I (Stefanakis, 2007; Parimala et al., 2011). DBScan is a density-based clustering algorithm, which does not need any pre-distribution knowledge. It can filter noise I while detecting CISs. Due to its goal, that is, to cluster insertions correctly, it can identify CIS boundaries maximally. In addition, its parameter (Eps) can meet any CIS width which is significant biologically. When detecting CISs by DBScan, all the insertions related to cancer genes are divided into core insertions and border insertions. Core insertions indicate those in CISs. Border insertions represent hotspots. So DBScan can resolve the issue of insertion biases. Secondly, based on standard deviation, we have developed NSD (normal standard deviation) to calculate DW to analyze how the insertions are distributed across tumors. To further our research step-by-step, we use tumor types instead of tumors in this paper. NSD overcomes two drawbacks of standard deviation in CIS detection: (1) standard deviation cannot compare two CIS involving different tumor types; and (2) standard deviation cannot compare two CISs with different insertion quantities. Due to the