International Research Journal of Biochemistry and Bioinformatics Vol. 8(2) pp 7-11, October, 2018
DOI: http:/dx.doi.org/10.14303/irjbb.2018.002
Available online http://www.interesjournals.org/IRJBB

Mini Review

Solid K-mer mining by using Z-score

Long Li, Liang Zhao*

School of Computing and Electronic Information, Guangxi University, Guangxi 530004, China
s080011@e.ntu.edu.sg

ABSTRACT

K-mers are fundamental building block for many NGS applications. However, k-mers are error prone, posing great challenges for downstream data analyses. We propose a statistical approach to effectively distinguish solid k-mers from weak k-mers. Precisely, we calculate a z-score for each k-mer, and jointly determine whether it is really solid based on its z-score and frequency. Experiments show that our approach effectively pinpoints out solid k-mers having low frequency, achieving an average improvement of 11.25%.

Keywords: K-mer, Z-score, next-generation sequence.

INTRODUCTION

The ever-increasing high throughput and dramatic decreasing cost for next-generation sequencing (NGS) technology provides great chance to revolutionize a wide range of medical and biological research as well as their induced application fields, such as medical diagnosis, biotechnologies, virology, etc (Alic et al., 2016; Zhao et al., 2017). However, the sequences are not perfect, and there exists various kind of errors, such as substituions, insertions, deletions, and uncalled bases, e.g., substituation error rates range from 1% to 2.5% and insertion and deletion error rate is as high as 40% (Kelley et al., 2010; Goodwin et al., 2015). The errors in a sequencing data has posed great challenges for data analysis. Hence, correcting these errors is the very first and critical step. Many downstream applications can be beneficial from corrected sequencing reads, such as sequence assembly, variants calling, reads mapping, etc (Salmela and Schroder, 2011). Dozens of approaches have been proposed to correct errors, just to mention a few, Coral (Salmela and Schroder, 2011), BLESS (Heo et al., 2014), MEC (Zhao et al., 2017). These approaches are heavily k-mer dependent.

A k-mer is a substring of a sequencing read having k consecutive bases. The very first and essential step of k-mer-based approach is usually mining of solid k-mers. A solid k-mer is considered as the one having frequency larger than a minimum threshold, while the rest are weak (Heo et al., 2014). Although this simple definition is effectively useful to distinguish solid k-mers from weak k-mers, it still has obvious limitations. The most important weakness is that a k-mer having low frequency may not be weak. This is because the sequencing depth is not uniform distributed due to system bias, e.g., GC-content regions are difficult to sequence. To this end, we focus on an important but less study problem of refining solid k-mers by using statistical analysis.

our model starts with counting k-mers by using KMC2 (Deorowicz et al., 2015), and splits all k-mers into solid and weak set tentatively based on their frequency. Later, z-score is computed for each k-mer and its solidity is determined based on the z-score as well as its frequency jointly.

MATERIALS AND METHODS

Data Sets

Four real NGS datasets collected in GAGE (Salzberg et al., 2011) is used to exploit the capability of our model in distcting solid k-mers from weak k-mers, i.e., Staphylococcus aureus (D1), Rhodobacter sphaeroides (D2), Human Chromosome 14 (D3) and Bombus impatiens (D4). We refer readers to GAGE for more details.

Pinpoint Minimum k-mer Frequency

KMC2 (Deorowicz et al., 2015) is borrowed in this study to count the frequencies of k-mers. Based on the results, we determine the minimum frequency used to separate weak k-mers from solid k-mers. Other than using various statistical models, such as Gamma + Gaussian + Zeta distribution based (Kelley et al., 2010), multinormial distribution based (Yang, et al. 2011), we simply set the minimum solid frequency of 5. That is, a k-mer is considered as weak tentatively if its frequency is less than 5. The rational is that an erroneous k-mer appears exactly 5 times is very small. Taking a human genome as an example, suppose reasonably that the whole genome length is 3×10^9, the sequencing depth is 30 and the error rate is 1%, then the number of erroneous k-mers appear exactly n times is $(1/3 \times 0.01)^n \times 30 \times 3 \times 10^9$. When n is 4, the value is 11.1; when n is 5, the value is 0.037.

Calculate z-Score of a k-mer

Given a k-mer κ, we define its neighbor as $N(\kappa)$ by $N(\kappa) = \{\kappa' : D(\kappa, \kappa') \le d_0, \kappa' \in K\}$, where $D(\kappa, \kappa')$ is the edit distance between κ and κ', and d_0 is the predefined maximum distance. The default value of d_0 is 1 as used in this study, but user can adjust this value to any reasonable integer. The k-mer cluster centered at κ is defined as $C(\kappa) = \{\kappa\} \cup N(\kappa)$, and the set of frequencies associated with these k-mers is defined as $F(\kappa) = \{f(\kappa) : \kappa \in C(\kappa)\}$. Based on $F(\kappa)$, the z-score of κ, $z(\kappa)$, is computed by $z(\kappa) = \frac{f(\kappa) - \mu}{\sigma}$, where μ is the averaged frequency of $F(\kappa)$ and σ is the standard deviation of $F(\kappa)$.

Refine Solid k-mers

A z-score of a k-mer as well as its frequency are jointly used to refine solid k-mers through the following two criteria: (i) if $f(\kappa) < f_0$ and $z(\kappa) \ge z_0$, then κ is moved from the weak k-mer set to the solid k-mer set and; (ii) if $f(\kappa) \ge f_0$ and $z(\kappa) < -z_0$, then κ is moved from the solid k-mer set to the weak k-mer set. The f_0 is the minimum frequency which is set as 5 in this study, and z_0 is the maximum z-score used to distinguish weak k-mers and solid k-mers. z_0 is learned from the z-score distribution of input data automatically. In this study, it is optimized to 0.8.

RESULTS

We conducted experiments on the four real data sets. Results show that z-score is able to refine both weak k-mers and solid k-mers, particularly useful for pinpointing out solid k-mers from weak k-mers, i.e., k-mers having low frequency but is correct in reality. An example of z-score distribution pertaining to k-mer frequency is shown in Figure 1, which is derived from Bombus impatiens at k=25. The highlighted k-mers shown in the figure have relatively low frequency—less than 8, while the z-score is pretty high—greater than 1. Interestingly, almost all solid k-mers (the top right region) have the similar level of z-score comparing to these highlighted ones. These observations indicate that the highlighted k-mers are very likely to be correct k-mers instead of erroneous ones although their frequency is very low. Hence, we move these k-mers from the weak set to solid set. The z-score distribution pertaining to the other three real data sets has similar patterns compared to the one shown here.

By exploring the four real data sets, we found that the proportion of k-mers that can be refined comparing to the solely frequency determined k-mers are 12.3%, 14.2%, 11.4%, 7.1% for D1, D2, D3 and D4, respectively. These refinements improve the purity of both weak k-mers and solid k-mers, which can be used for error correction in the downstream data analyses.

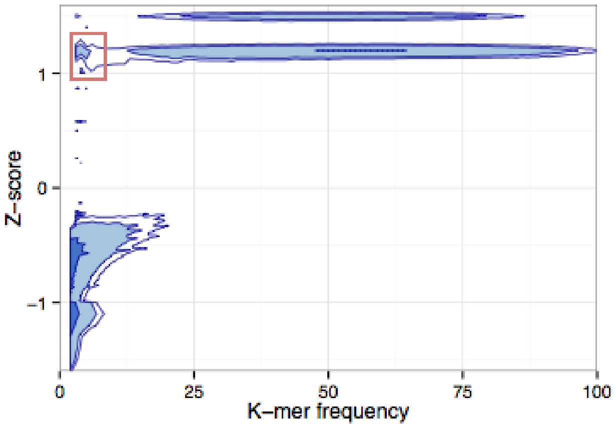

Figure 1. The relation between z-score and k-mer frequency.

The level of shade represents the density of the distribution. The darker the color is, the more k-mers are presented. The frequency of the k-mers highlighted in the red box are less than nine, which are very likely to be treated as weak for all existing k-mer based approaches. However, they should be considered as solid K-mers based on the very high z-score they have. The data shown here is obtained from B. impatiens at K=25.

CONCLUSION

A k-mer is a fundamental building block for many sequencing analysis, particularly useful for error correcton, sequence assembly, variants calling etc. However, due to sequencing errors and bias, a k-mer having low frequency may not be erroneous. Hence, the distinguishing of solid k-mers only by frequency is not optimal. Instead of overlook this issue by existing approaches, we propose a novel idea of using z-score to distinguish erroneously classified weak and solid k-mers. Experiments show that z-score is sufficiently useful to distinguish real solid k-mers. The average proportion of refined k-mers is 11.25% for the four real data sets.

REFERENCES

Alic AS, Blanquer I , Dopazo J and Ruzafa D (2016). Objective review of de novo stand-alone error correction methods for NGS data. WIREs. Comput. Mol. Sci. 6(2): 111-146.

Aluru S, Dorman KS and Yang X (2011). Repeat-aware modeling and correction of short read errors. BMC. Bioinformatics. 12(1) : S52.

Chen D, Hwu WM, Heo Y, Ma J and Wu XL (2014). BLESS: Bloom filter-based error correction solution for high-throughput sequencing reads. Bioinformatics, 30(10): 1354-1362.

Chen Q, Jiang P, Li W, Li J, Wong L and Zhao L (2017). MapReduce for accurate error correction of next generation sequencing data. Bioinformatics. 33(23): 3844-3851.

Delcher AL, Koren S, Magoc T, Marcais G, Puiu D, Pop M, Phillippy AM, Roberts M, Salzberg SL, Schatz MC, Treangen TJ, Yorke JA and Zimin A (2011). GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome. Res. 22(3) : 557-567.

Deorowicz S, Debudaj-Grabysz A, Grabowski S and Kokot M (2015). KMC 2: fast and resource-frugal k-mer counting. Bioinformatics. 31(10): 1569-1576.

Deshpande P, Ethe-Sayers S, Goodwin S, Gurtowski J, McCombie WR and Schatz MC (2015). Oxford nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome. Res. 25: 1750-1756.

Kelley DR, Schatz MC, Salzberg SL (2010). Quake: Quality-aware detection and correction of sequencing errors. Genome. Biol. 11(11) : R116.

Salmela L, Schroder J (2011). Correcting errors in short reads by multiple alignments. Bioinformatics. 27(11) : 1455-1461.