2018 2nd International Conference on Modeling, Simulation and Optimization Technologies and Applications (MSOTA 2018)
ISBN: 978-1-60595-594-0

A General Overview of Detection Methods for Single Nucleotide Polymorphism

Hai YANG

School of Information Science and Electrical Engineering, Shandong Jiaotong University, Jinan, Shandong, China

Keywords: Single Nucleotide Polymorphism, Next Generation Sequencing, HMM, MAQ.

Abstract. Single nucleotide polymorphism, often abbreviated to SNP, in the genomes of species individuals refers to a variation in a single nucleotide that occurs at a specific position in the genome, where each variation is present to some appreciable degree within a population. In recent years, the research of detection algorithm for SNP has been a hotspot in genomics, computational biology and informatics. Most of SNP detection algorithms are based on sequencing technology of the next generation sequencing (NGS). In this paper, a general overview of detection methods for single nucleotide polymorphism is made to deeply analyze several current mainstream detection algorithms for SNP. It is important for further development of this field of bioinformatics.

Introduction

In recent years, with the development of bioinformatics and molecular detection technology, the research of single nucleotide polymorphism (SNP) has attracted much attention all over the world [1]. Generally speaking, SNP refers to a DNA sequence polymorphism caused by single-nucleotide variation at the genome level. For a single individual, SNP refers to polymorphic sites in the genome. For the population, SNP can be regarded as the base position with polymorphism in the population [2]. As a genetic variation, SNP can describe the vast majority of differences between individuals and groups of organisms [3]. Therefore, the SNP detection has been widely used in genomic research.

The basic processes of SNP detection include genotype resequencing, sequence alignment, SNP detection and filtering of results [4]. Since the vast majority of SNP sites are located in highly interlocking and unbalanced genomic regions, a representative group of SNP sites (Haplotype, known as Haplotype) can be selected as the research object [5], which helps to identify similar polymorphic sites near chromosomes. Due to the importance of SNP sites in biogenetics, the purpose of this paper is to overview the current mainstream detection algorithms and research progress of SNP in academia.

Sequence Alignment

Nowadays, the DNA sequencing technology has gone through four generations. The first generation of DNA sequence analysis method is based on accurate sequence or genotype [6]. The second generation sequencing technology which is known as high-throughput sequencing technologies, depends on the mass marking nucleotide. The third generation sequencing technology is based on direct observations of DNA replication, and the fourth generation of sequencing technology is based on nanopores single DNA sequencing technology. Compared with the classical first-generation sequencing, the second-generation sequencing technology has the advantages of low experimental cost, low experimental complexity and high output rate. With the development of various algorithms, it has been widely used in biological studies such as genomic DNA sequencing, genetic modification and RNA transcriptome [7].

Sequence alignment is to compare the short sequence obtained by resequencing in contrast to the existing genome sequence (reference) and precisely locate the position of the short sequence on the

reference genome sequence [8]. Through sequence comparison, we can find a lot of biological information related to the genetic evolution of organisms.

The main idea of two-sequence alignment is based on the similarity between the two sequences. At present, needleman-wunsch dynamic programming algorithm is a relatively effective method for global sequence alignment, while smith-waterman algorithm which has made some improvements on the former is applicable to local alignment. However, the double-sequence alignment is not applicable to second-generation sequencing data with high throughput. In order to cope with the development of sequencing technology, the introduction of heuristic algorithm is particularly necessary [9]. At present, most heuristic algorithms speed up the computation by means of indexing. The two commonly used data structures are hash table and suffix tree based on BWT. The characteristics of hashing table are that it is more sensitive to the matching results with high accuracy and sensitivity to the detection of SNP sites. The suffix tree can filter the matching sequences with low confidence value and avoid some unnecessary work in the process of comparison and analysis.

The Detection Algorithm for SNP

The genome data of an organism, which includes the mass fraction of each base, actually indicates the probability that the base will be misidentified during sequencing. The Phred-base mass fraction formula is usually used to determine the base error rate, which can be calculated by formula (1) as follows:

$$Q_{-score} = -10 \times \log_{10} P \tag{1}$$

The sequence alignment process can obtain a series of statistics, including the total number of aligned reads, number of mismatches, gap size, etc., which can be used to evaluate the quality of data and calibrate its correctness. When filtering the results of alignment, the sequences with the alignment quality less than 20 or the matching value greater than 2 can be directly filtered out to obtain more accurate data. After the above series of data preprocessing, we can use various snp-calling detection algorithms to excavate single nucleotide polymorphisms in the genomes of organisms. The current mainstream SNP detection algorithms include MAQ, Samtools and haplotype detection algorithm, etc.

Core Algorithm of MAQ

MAQ, on behalf of mapping and assembly with qualities, is a diversity filtering algorithm based on short sequences alignment and can be used to address very short sequence segments generated by NGS technology. MAQ can quickly align short sequences, rectify the process by comparing with paired information, and estimate consistent sequences by using bayesian model. The biggest advantage of MAQ algorithm is that it requires less memory to perform snp detection of a single individual, which is faster and more convenient than other algorithms. However, MAQ algorithm is mainly aimed at a single individual. When there are hundreds of individuals for comparison, the algorithm takes serious time. MAQ algorithm is not suitable for the alignment of long sequences. MAQ algorithm assumes that the research object is a diploid organism, and the observed data set is D, where the main base is $<b,b^{'}>\in\{A,C,G,T\}$. Since SNP variation is usually second-order, the error probability of n records (Reads) detected at a certain site of diploid organisms is:

$$\varepsilon_1 \le \varepsilon_2 \cdots \le \varepsilon_k \tag{2}$$

$$\varepsilon_{k+1} \le \varepsilon_{k+2} \cdots \le \varepsilon_n \tag{3}$$

When the genotype is heterozygous, the influence caused by errors in sequencing and alignment can be ignored if the mass fraction of the base is high enough. The binomial distribution model with a parameter of 0.5 can be directly used. The likelihood function is as follows:

$$P(D|<b,b^1>) = \binom{n}{k} / 2^n \quad (4)$$

The probability of exactly k errors in n bases is defined as a_{nk}, which can be calculated by following formula:

$$P(D|<b,b>) = a_{nk} \quad (5)$$

When the measurement has the same error probability which is independent and consistent, the $\bar{a}_{nk}(\bar{\varepsilon})$ can be calculated by binomial distribution, described as formula (6):

$$\bar{a}_{nk}(\bar{\varepsilon}) = \binom{n}{k} \bar{\varepsilon}_k (1-\bar{\varepsilon})^{n-k} \quad (6)$$

When the error probability is correlated and not identical, it can be approximate to:

$$a_{nk} = c'_{nk} \prod_{i=0}^{k-1} \varepsilon_{i+1}^{\theta^i} \quad (7)$$

Where ε_i is the minimum base error probability, c'_{nk} is a function of ε_i which is not sensitive to the changes of ε_i. As a result of larger experimental data, $\theta = 0.85$.

Next, let r be the observation probability of heterozygosity. The prior probability of genotype can by calculated by these formulas:

$$P(<b,b^1>) = r \quad (8)$$

$$P(<b,b>) = P(<b,b^1>) = \frac{1-r}{2} \quad (9)$$

By calculating the posterior probability P(g|D) of genotype g (given observation value D), genotype and mass fraction can be estimated:

$$\hat{g} = \arg\max_g P(g \mid D) \quad (10)$$

$$Q_g = -10 \times \log_{10}[1 - P(\hat{g} \mid D)] \quad (11)$$

Usually the parameters $r = 0.001$ is set to detect SNP and $r = 0.2$ is set to infer the genotype of known SNP sites.

Core Algorithm of Samtools

The core components of Samtools include Samtools and Bcftools. Samtools is used to improve the error-detection probability to infer likelihood functions, while Bcftools is used to do the snp-calling and some statistics estimation. Samtools is an optimization on the MAQ model, which is applicable to multiple samples with low coverage, which can make up for the disadvantage of MAQ algorithm only suitable for defection of a single individual. Moreover the running speed of Samtools algorithm is relatively fast. However the quality value of variation loci obtained by Samtools is relatively low because the algorithm is a bit of simple.

Similarly, under the condition that the observation data set is D, Samtools algorithm assumed that there were n reads at a certain site for a single sample. The bases of first l reads are the same with those in the reference genome, and the bases of rest $n-l$ reads are not same with those of reference genome. The error-detection probability of the j-th read is ε_j. Therefore the likelihood function of genotype g can be described as formula (12):

$$L(g) = \frac{1}{m^k} \prod_{j=1}^{l} [(m-g)\varepsilon_j + g(1-\varepsilon_j)] \prod_{j=l+1}^{n} [(m-g)(1-\varepsilon_j) + g\varepsilon_j] \quad (12)$$

Where m is the ploidy of the species individual. For the multiple samples, under the case of observation data is D, the probability of loci i is a polymorphism loci is g_i. Its haploid type is h_i, ploidy is m_i, so the genotypes can be estimated by the bayes principle as follows:

$$\hat{g} = \arg\max_{gi} \frac{P(D_i \mid g_i)P(g_i \mid \Phi_E)}{\sum_{hi} P(D_i \mid h_i)P(h_i \mid \Phi_E)} \quad (13)$$

$$\Phi_E = \frac{E(X \mid D, \Phi)}{M} \quad (14)$$

Where X is the whole genome, $M = \sum_i m_i$, the genotype frequency of loci is Φ_E, which is optimized using the EM algorithm.

The Haplotype Detection Algorithm

Haplotype detection is a calling method that could obtain the SNP and Indel (insertion and deletion) simultaneously through local recombination of active regions. It is one of the main models of GATK, which is one of the most mainstream software.

The advantage of Haplotype detection lies in the introduction of real-time Denovo algorithm, which greatly reduces false positives caused by alignment errors, especially in Indel detection [10]. However, haplotype detection algorithm also has its own disadvantages, that is, it runs very slowly. In addition, it doesn't support the detection process only for SNP or Indel.

The Haplotype detection is divided into the following four steps: firstly define the active area, and then determine the restructuring of the active region may haploid type; thirdly calculate the probability of each haplotype according to the given reads information; finally calculate the posteriori probability of each haplotype using the bayesian formula so as to select the genotype with the largest probability as the genotype of sample.

The Calculation of Haplotype Using Pair HMM

The pair hidden markov model (Pair HMM) is an important tool for the Haplotype Detection Algorithm. After obtaining the possible haplotypes of organisms, we need to evaluate the reliability of their different haplotypes. In haplotype detection, Pair HMM algorithm is usually used in this step to rematch sequences (Figure 1), and the probability of each haplotype is obtained by combining its information of data quality.

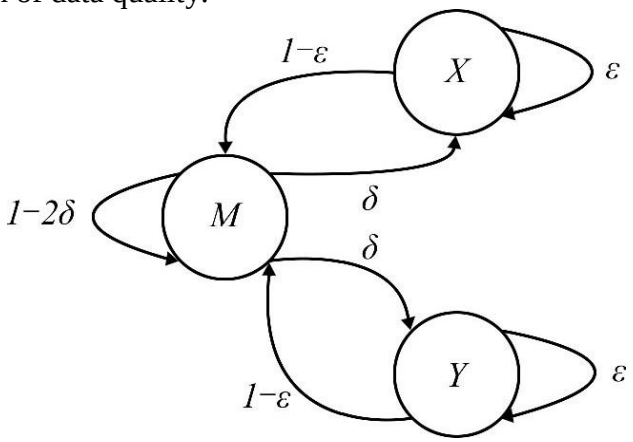

Figure 1. The flowchart of Pair HMM.

For sequence alignment, there are two hidden states, base matching and Indel, and the M, X, and Y states correspond to base matching, base deletion, and base insertion (Table 1). The probability of transition δ denotes the indel appearing and the probability of transition ε denotes the indel maintaining respectively.

Table 1. Corresponding relations among the hidden states of Pair HMM.

	M	X	Y
M	$1-2\delta$	δ	δ
X	$1-\varepsilon$	ε	0
Y	$1-\varepsilon$	0	ε

The probability of output for each hidden state is given by the mass score of sequencing alignment. Then we get the iterative formula as follows:

$$P_M = P_{xi,yj} \max\begin{bmatrix}(1-2\delta)P_M(i-1,j-1)\\(1-\varepsilon)P_X(i-1,j-1)\\(1-\varepsilon)P_Y(i-1,j-1)\end{bmatrix} \tag{15}$$

$$P_X(i,j) = q_{xi} \max\begin{pmatrix}\delta P_M(i-1,j)\\\varepsilon P_X(i-1,j)\end{pmatrix} \tag{16}$$

$$P_Y(i,j) = q_{yj} \max\begin{pmatrix}\delta P_M(i,j-1)\\\varepsilon P_Y(i,j-1)\end{pmatrix} \tag{17}$$

Through the above iteration, we can obtain a probability matrix with a given reads as rows and haplotype as columns:

$$\begin{bmatrix}A_{11} & A_{12} & \cdots & A_{1m}\\A_{21} & A_{22} & \cdots & A_{2m}\\\vdots & \vdots & \vdots & \vdots\\A_{n1} & A_{n2} & \cdots & A_{nm}\end{bmatrix} \tag{18}$$

Therefore, the likelihood value of a single loci allele of a given read is given by the maximum probability corresponding to the haplotype containing the allele.

Based on bayesian full probability formula, under the condition of observed data D, the posterior probability of genotype G is:

$$P(G \mid D) = \frac{P(G)P(D \mid G)}{\sum_i P(G_i)P(D \mid G_i)} \tag{19}$$

$$P(D \mid G) = \prod_j \left(\frac{P(D_j \mid H_1)}{2} + \frac{P(D_j \mid H_2)}{2}\right) \tag{20}$$

So, we can get the posterior probability of every possible base type at each loci, under the given observations. We define the genotype with the highest probability as the genotype of this loci, and give the confidence level of SNP for the corresponding loci.

Conclusion

With the continuous development of research technology, modern SNP technology plays an increasingly important role in the genetic factor analysis of complex diseases. At the same time, SNP technology is also of great value for the analysis of plant genetic variation, and has broad

application prospects in the analysis of population structure, map construction, quantitative trait locus location, association mapping, mapping cloning and marker-assisted selection. Although the technology of SNP detection and detection in diploid organisms have been constantly optimized and improved, the detection of heteropolyploid SNPs still has some limitations. In fact, many important crops in nature are allopolyploids, and the sequence variation between their subgenomes coexists with the allele variation in the subgenome. Therefore, there is a high level of genomic complexity in allopolyploid crops, which poses a great challenge to their SNP detection technology.

Recently, although the development of SNP in polyploid plants is slow, due to the development of high-throughput sequencing technology and continuous optimization of SNP detection methods, it is gradually possible to explore SNP in heteropolyploid plants, and it is expected to provide effective means for genetic mapping and correlation analysis of polyploid plants.

Reference

[1] Yu X., and Sun S., Comparing a few SNP calling algorithms using low-coverage sequencing data, BMC Bioinformatics, 2013, 14 (1) :274.

[2] Wang H., Liu J., Fu L., and Mei D.S., Review on single nucleotide in polymorphisms in polyploid crop Brassica napus. Chinese Journal of Oil Crop Sciences, 2014, 36(3): 422-429.

[3] Xu J.L., Wang Y., Hou M., and Li Q., Progression detection methods of SNP. Molecular Plant Breeding, 2015, 13(2): 475-482.

[4] Zhang Y., and Wang R., Sequence alignment algorithms in bioinformatics, Computer Knowledge and Technology, 2008, 1(1): 181-184.

[5] He D.H., Xing H.Y., Zhao J.X., Zhao Y.N., Qi C., and Wang YT., Single nucleotide polymorphism SNP discovery in polyploidy plants, Journal of Zhejiang University (Agriculture and Life Science), 2011, 37(5): 485-492.

[6] Nielsen R., Korneliussen T., Albrechtsen A., Li Y., and Wang J., SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data PLoS One, 2012, 7(7): e37558.

[7] Hodgkinson A., and Eyre-Walker A., Human triallelic sites: evidence for a new mutational mechanism, Genetics, 2010, 184(1):233-241.

[8] Jiao Y., Gao J., and Zhang W.G., Research progress in pairwise sequence alignment algorithms and their software, Computer Application and Software, 2015, 32(6):5-8.

[9] Mielczarek M., and Szyda J., Review of alignment and SNP calling algorithms for next-generation sequencing data J. Appl. Genet., 2015, 57(1):71-79.

[10] Li J., Pan Y.C., Li Y.X., and Shi T.L., Analysis and application of SNP and haplotype in the human genome. Acta Genetica Sinica, 2005, 32(8): 879-889.