Repeat finding by normal approximation on whole genome shotgun assembling

Naotoshi Seo s1080134 Supervised by Hiroshi Toyoizumi

Abstract

This paper proposes a method to find repeat by normal approximation on a whole genome shotgun assembling. We stochastically modeled the redundancy of subsequences with same arrangement. Traditionally, it is modeled by a Poisson distribution. However, this model was actually a binomial distribution, and a binomial distribution resembled a normal distribution than a Poisson distribution. We showed that a normal approximation is more efficient than the traditional Poisson approximation and proposed a method to estimate the most effective threshold value for finding repeat. And then we verified it by experiments using our simulator programs.

1 Introduction

1.1 Basic knowledge about genome

All creatures consist of cells, and each cell possess genetic information to produce protein. This genetic information is preserved in chromosomes that exist in the core of a cell. For example, a human has 46 chromosomes (2 sets of 22 autosomal chromosomes, two sex chromosomes), and all genetic codes exist in the chromosomes. All these chromosomes are collectively called a genome. That is, genome means whole genetic information. Chromosomes consist of DNA (Deoxyribonucleic Acid). A DNA molecule has a twisted spiral form called a double helix. A DNA consists of four nucleotides: A (Adenine), T (Thymine), G (Guanine), and C (Cytosine). Because A combines with only T, and C combines with only G, the one base sequence of double helix will also determine another sequence. Three nucleotides comprise an amino acid and these amino acids are information to generate proteins.

1.2 Whole genome shotgun assembling

It is impossible to read a genome at a burst by current technology because the DNA of a human is too long with three billion characters. Therefore, a technology which reassembles fragments after cutting DNA to a length that can be read is needed. The technology is DNA fragment assembling. There are well known techniques such as the whole genome shotgun assembling and the hierarchical shotgun assembling. The International Human Genome Project team has used the hierarchical shotgun assembling; in contrast, the Celera Genomics corporation has used the whole genome shotgun assembling to decipher the human genome [9].

In whole genome shotgun assembling, the genome DNA is first copied and increased. Second, these DNA sequences are cut to small fragments by several restriction enzymes. Third, the base sequences of the both ends of fragments are read. Fourth, the whole sequences are linked up by the subtle overlaps. At last, the original sequences are recomposed. The whole genome shotgun assembling is superior to the hierarchical shotgun assembling in regard to efficiency and speed; however, it is inferior in regard to precision. It is theoretically supposed that ten copies are enough. The Celera Genomics corporation finished reading the whole base sequences of human DNA in June 2000 using this method [1].

1.3 The purpose of this research

Repeat means subsequences with the same arrangement in one genome sequence. A repeat must not use for over-

——— GCAGTACG ——— GCAGTACG ———-

Figure 1: Repeat example

lap detection because its original location can not be determined by shotgun assembling. Therefore, methods for finding repeat are required. It can be judged whether the subsequence is repeat or not by the redundancy of subsequences with same arrangement. Traditionally, the distribution of it is estimated as a Poisson distribution [4] and the threshold value for finding repeat is estimated in the model. However, it is doubtful that the traditional method is valid. In this paper, the normal approximation is proposed as a new method and the higher efficiency of it is shown. Furthermore, a method to estimate the efficient threshold value for repeat finding is proposed. These are estimated stochastically and verified by experiments using our simulator program.

2 Estimation Technology

In this paper, after first estimating the stochastic model, its validation *was checked* by a simulator. The estimation *was produced* mathematically and *was calculated* by **Mathematica** [12]. The simulator *used* an algorithm with whole genome shotgun assembling, and it *referred*

the algorithm of Celera Genomics corporation [3]. A detailed description about the algorithm is given in Section 5, Appendix. The simulator *used* a few genome data of the HIV virus and E.coli as typical prokaryotic genomes. The genome sequences files *were downloaded* from **GenoBase** [8] and **HIV Sequence Database** [5]. In details, the simulator *were created* by **ruby language** [2] which *used* a **mysql database** [7] and **ruby-mysql** [6], a connector program between ruby and mysql.

3 Repeat Finding

One genome sequence has subsequences with the same arrangement called *repeat* in the different parts. Therefore, this makes it difficult to differentiate one fragment from other fragments and to know where these fragments are originally located. Therefore, repeats must not used for overlap detection for reassembling. A threshold value is needed to judge fragments as a repeat. The fragments are judged as a repeat if the # of copies is larger than the threshold. In simple thinking, when a genome sequence is increased to a number such as 3, the redundancy of one subsequence ordinarily becomes 3. If the genome has another subsequence with its same arrangement, in short, *repeat*, the number becomes 6. Actually, these numbers decrease and the threshold becomes obscure because DNA is fragmented and all subsequences do not appear and a machine may fail reading the fragments. Thus, we need to use the probability theorem and estimate it.

3.1 Estimation of the redundancy

Here we will use the following definitions:

n = the # of copies,
W = word (subsequence) length,
N = the redundancy of subsequences,
p_c = cut probability, and
p_m = miss reading probability.

The cut probability p_c is a probability that restriction enzymes cut genome sequences. The miss reading probability p_m is a probability that a machine miss reading fragments. The probability of not being cut in length W, in other words, the probability with complete subsequence,

$$q = P\{W = w\} = (1 - p_c)^w. \tag{1}$$

The probability that the # of the complete subsequence is y when the # of copies is n can be obtained by combination of y in n with probability q. This means that the distribution of it takes a binomial distribution. Therefore, the probability becomes

$$ {}_nC_y q^y (1-q)^{(n-y)}. \tag{2}$$

We set a probability r that a sequence is correctly read by machine,

$$r = 1 - p_m \tag{3}$$

The distribution of the number x correctly read by machine from y number also takes a binomial distribution. The probability becomes

$$ {}_yC_x r^x (1-r)^{(n-x)}. \tag{4}$$

Therefore, the probability that the redundancy of subsequences with the same arrangement is N becomes

$$\begin{aligned} P\{N = x\} &= \sum_{y=x}^{n} \left(\begin{array}{c} n \\ y \end{array} \right) q^y (1-q)^{(n-y)} \left(\begin{array}{c} y \\ x \end{array} \right) r^x (1-r)^{(y-x)} \\ &= \left(\begin{array}{c} n \\ x \end{array} \right) (qr)^x (1-qr)^{(n-x)}. \end{aligned} \tag{5}$$

This is also a binomial distribution.

When the # of copy n was set to 10, word length w was set to 100, and cut probability p_c was set to $1/500$, this estimation results will be shown in Figure 2, and the simulator results will be shown in Figure 3. It seems that our estimation is correct.

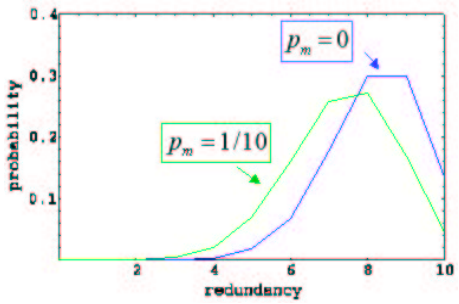

Figure 2: Estimation

3.2 A problem of the traditional approximation

A binomial distribution requires much time for calculation. Therefore, it is better to approximate to another distribution. Recently, the distribution of the number of fragments with the same arrangement is approximated by a Poisson distribution [4],

$$P\{N = x\} = e^{-\lambda} \frac{\lambda^x}{x!}, \tag{6}$$

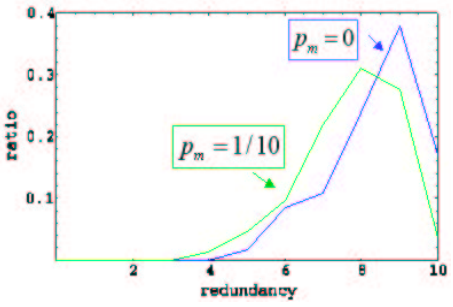

Figure 3: Simulator result

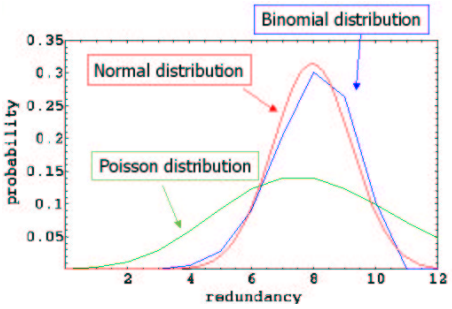

Figure 4: Comparison with other distributions

where N is the # of redundancy. Although a binomial distribution approximates to a Poisson distribution when n is sufficiently large, p is sufficiently small, and λ is set to $n \cdot p$, it does not approximate to a Poisson distribution because the number of copies is of a low value such as 10 and the probability p is large such as 0.8 in this case. The distribution does not approximate to a Poisson distribution even if n becomes large when p is large as this. It is needed to enlarge cut probability p_c or miss reading probability p_m to minify p. However, it is better that these parameters are as small as possible because reassembling would fail if the fragments' length is very small and the machine can not read fragments well if p_m is large. First of all, it is better not to increase the number of copies because the smaller it is, the shorter the time of copying the DNA in wetlab and reassembling in computer becomes. From above consideration, the traditional approximation would not be realized more if the performance of machines is improved in the future.

3.3 Prescription by normal approximation

A binomial distribution can be approximated by not only a Poisson distribution but also a normal distribution when n is sufficiently large or $p \doteq 0.5$. Although approximation is impossible because n is small such as 10 in this case like a Poisson approximation, resemblance may be possible than one by a Poisson distribution at least. Figure 4 shows that a binomial distribution resembles a normal distribution rather than a Poisson distribution. Even if copy number n, cut probability p_c, and miss reading probability p_m are changed, this result is unchanged. This result becomes more prominent if p_c and p_m get smaller by improvement of the machines' performance. The probability density function of a normal distribution is as follows,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right). \quad (7)$$

The variables, population mean u and population variance σ^2 are set as the following,

$$\mu = n \cdot p, \ and \quad (8)$$
$$\sigma^2 = n \cdot p \cdot (1-p) \quad (9)$$

because the mean of a binomial distribution is $n \cdot p$ and the variance is $n \cdot p \cdot (1-p)$. Normal approximation probably yields better results than the traditional Poisson approximation. Here, p is set to qr in section 3.1 for approximating a binomial distribution to a normal distribution.

3.4 The necessary copy number for repeat finding

Figure 5 shows that big error will occur for differentiating whether a word is repeat or not if the copy number is small whatever threshold will be taken. Figure 6 shows that the error become small if the copy number is big. In these figures, the right peak shows the distribution of the redundancy of subsequences having double repeat, and the left peak shows that one having no repeat. Copies more than certain number are required for valid repeat finding.

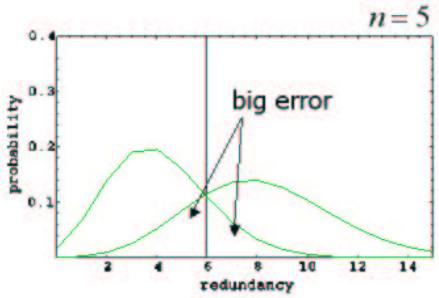

Figure 5: Big error

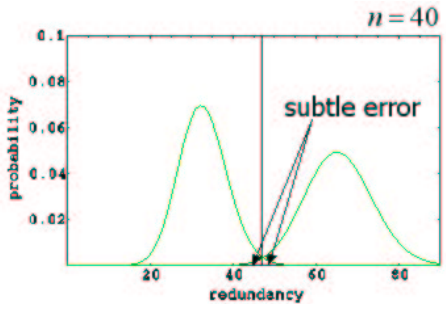

Figure 6: Subtle error

cut probability p_c	miss reading probability p_m	number of copies n	threshold value	false negative ratio	false positive ratio
1/500	0	5	5.8	0.0	0.0
1/500	1/10	8	8.6	0.0	0.0
1/500	1/5	12	11.4	0.0	0.0196
1/200	0	15	13.1	0.0	0.0035
1/200	1/2	52	30.2	0.0	0.0208

Table 1: Repeat finding experiments: word length = 100

We assumed that the error is permitted if $false\ positive\ ratio + false\ negative\ ratio < 0.05$. False positive means what is judged to be positive although it is not positive in fact. In this case, positive is repeat. False negative is the opposite, what is judged to be negative although it is not negative in fact.

When $p_c = 1/500$, and $p_m = 0$, the error ratio became less than 0.05 when n is 4 on a binomial distribution if good threshold value is used. It did when n is 5 in a normal approximation. It did when n is 28 in a Poisson approximation. In this case, about 1/6 copies are enough in a normal approximation compared with a traditional Poisson approximation. In other words, more effective estimation can be performed by a normal approximation than a traditional Poisson approximation.

3.5 The most effective threshold value

The most effective threshold value is the intersection's x-coordinate of a no-repeat distribution's curve and a double-repeat distribution's curve. The false negative of triple or more than repeat distributions is not necessary to count because it is very small compared with that of double-repeat distribution. In other words, the solution x of

$$\frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left(\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right) = \frac{1}{\sqrt{2\pi\sigma_2^2}} exp\left(\frac{-(x-\mu_2)^2}{2\sigma_2^2}\right) \tag{10}$$

is the most effective threshold value. Here,

$$\mu_1 = n \cdot p, \tag{11}$$
$$\sigma_1 = \sqrt{n \cdot p \cdot (1-p)}, \tag{12}$$
$$\mu_2 = 2n \cdot p, and \tag{13}$$
$$\sigma_2 = \sqrt{2n \cdot p \cdot (1-p)}. \tag{14}$$

3.5.1 Experimental proof

Table 1 shows that the last calculation for threshold makes good results. Good results mean the error ratio, $false\ positive\ ratio + false\ negative\ ratio$, is less than 0.05. The sufficient copy numbers were stochastically calculated so that at least one threshold can take error ratio less than 0.05. And then, the most effective threshold values were calculated by the above method.

Although good results occurred, better results would occur if the ratio of repeat was preexamined. It is why the actual distribution more resembles the distribution added double-repeat distribution multiplied by the ratio of repeat and no-repeat distribution multiplied by the ratio of not repeat. Figure 7 shows that the most effective threshold value would become more large when the ratio of repeat was less than that of not one like the sequence analyzed in above experiments. Better results could be

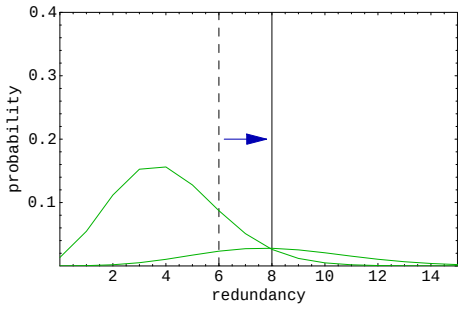

Figure 7: The distribution considered repeat ratio

achieved by Bayes statistical techniques using previous knowledges such as repeat ratio if repeat ratio is determined by species of creatures. However, this method is not deeply explored in this thesis because the purpose of this thesis is to propose an universal method to estimate the effective threshold value in whatever sequence is used.

4 Conclusion and Future Works

This paper has presented a method to find a *repeat* by normal approximation on whole genome shotgun assembling. It was showed that a binomial distribution resembled a normal distribution rather than a Poisson distribution. Therefore, normal approximation was used instead of traditional Poisson approximation. It was also showed that more effective estimation can be performed by a normal approximation than a traditional Poisson approximation. The most effective threshold value was calculated, and this threshold value indeed achieved good results in our simulations.

In this thesis, an universal method not based on kinds of DNA sequences was proposed to estimate the effective threshold value. Better results could be acquired by Bayes statistical techniques using previous knowledges because repeat ratio may be obtained if the kind of analyzed DNA sequences is determined such as of humans. Analysis using this Bayes statistical technique has a worth to be researched.

5 Appendix: Algorithm of Whole Genome Shotgun Assembling

First, the simulators *increased* the number of the genome data and *fragmented* in a random probability as the length of fragments have a exponential distribution [10]. Second, the simulators *reassembled* the fragments using an algorithm with whole genome shotgun assembling. We referred to the algorithm that Celera Genomics corporation used [3]. The algorithm for reassembling does the following:

1. Lists all k-long words in fragments and their reverse complements. This step takes advantage of the fact that we do not know which strand of DNA the fragment came from and the possibility of generating overlaps on the complementary strand.

2. Using that list of k-long words, a table is built with the following fields:
   WordID - FragmentID - Orientation - Position in Fragment.

3. Sorts the table by the WordID. This step allows us to quickly detect overlaps and also to apply heuristics for eliminating repeats.

4. Finds word matches in the table, which would indicate an overlap.

5. Drops WordIDs with too many matches, as this would indicate a probable repeat.

6. For each pair of fragments containing a common WordID, the common word is found and a narrow channel is built (based on the max error size) around the word to search using an alignment algorithm like Needleman-Wunsch [11]. A table is built with the following fields:
   FragmentID1 - Orientation1 - LeftPosition1 - RightPosition1 - FragmentID2 - Orientation2 - LeftPosition2 - RightPosition2.
   Length parameter is also possible instead of RightPosition1 and RightPosition2.

7. Finds fragments connected from one fragment in the table.

8. Connects them with paying attention to the orientation. In our simulator, a fragment most extended to a long distance was connected when there were two or more fragments connected from one fragment.

9. Forms islands by connecting fragments recursively.

Although the alignment algorithm is used for detecting the resemblance between two fragments with possible read errors in Celera Genomics cooperation, the alignment algorithm was not used in our simulator because fragments are matched 100%.

6 Acknowledgments

I would like to thank Prof. Hiroshi Toyoizumi for his advice and Prof. Stephen G. Lambacher for his help in improving my English writing.

References

[1] Celera, Celera Genomics: Press Release, 2000, http://www.celera.com/celera/pr_1056581295.

[2] Hiroyuki Matsumoto, The Object-Oriented Scripting Language Ruby, (current Nov. 2003), http://www.ruby-lang.org/ja/.

[3] M. F. Kim and R. May, "Cs262: Lecture 8 notes whole genome shotgun," http://www.stanford.edu/class/cs262/Notes/ln8.pdf.

[4] E. Lander and M. Waterman, "Genomic mapping by fingerprinting random clones: a mathematical analysis," Genomics, 2, 231-239, 1998, http://www-mig.jouy.inra.fr/stat/pagesperso/papiers/Sch95.ps.gz.

[5] Loa Alamos National Laboratory, HIV Sequence Database, (current Oct. 2003), http://www.hiv.lanl.gov/content/hiv-db/mainpage.html.

[6] Masahiro Tomita, MySQL - Ruby Interface, (current Nov. 2003), http://www.tmtm.org/mysql/ruby/.

[7] mysql.com, mysql, (current Nov. 2003), http://www.mysql.com.

[8] Nara Institute of Science and Technology, GenoBase, (current Nov. 2003), http://ecoli.aist-nara.ac.jp/gb4/common_data/ecoli-full.fasta.

[9] NATIONAL INSTITUTES OF HEALTH, NIH NEWS RELEASE, 2000, http://www.nih.gov/news/pr/jun2000/nhgri-26.htm.

[10] J. K. Percus, Mathematics of Genome Analysis, chap. 1. Decomposing DNA and 2. Recomposing DNA, CAMBRIDGE UNIVERSITY PRESS, 2002.

[11] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, Biological sequence analysis, chap. 2:19-20, CAMBRIDGE UNIVERSITY PRESS, 2000.

[12] WOLFRAM RESEARCH, Mathematica, (current Dec. 2003), http://www.wolfram.com/.