A Statistical Approach Designed for Finding Mathematically Defined Repeats in Shotgun Data and Determining the Length Distribution of Clone-Inserts

Lan Zhong[1, 2], Kunlin Zhang[1], Xiangang Huang[2], Peixiang Ni[2], Yujun Han[2], Kai Wang[2], Jun Wang[1,2], and Songgang Li[1, 2*]

[1] *College of Life Science, Peking University, Beijing 100871, China;* [2] *Beijing Genomics Institute/Center of Genomics and Bioinformatics, Chinese Academy of Sciences, Beijing 101300, China*

The large amount of repeats, especially high copy repeats, in the genomes of higher animals and plants makes whole genome assembly (WGA) quite difficult. In order to solve this problem, we tried to identify repeats and mask them prior to assembly even at the stage of genome survey. It is known that repeats of different copy number have different probabilities of appearance in shotgun data, so based on this principle, we constructed a statistical model and inferred criteria for mathematically defined repeats (MDRs) at different shotgun coverages. According to these criteria, we developed software MDRmasker to identify and mask MDRs in shotgun data. With repeats masked prior to assembly, the speed of assembly was increased with lower error probability. In addition, clone-insert size affects the accuracy of repeat assembly and scaffold construction. We also designed length distribution of clone-inserts using our model. In our simulated genomes of human and rice, the length distribution of repeats is different, so their optimal length distributions of clone-inserts were not the same. Thus with optimal length distribution of clone-inserts, a given genome could be assembled better at lower coverage.

Key words: mathematically defined repeat (MDR), clone-inserts, assembly

Introduction

Sanger invented the DNA sequencing technology using dideoxynucleotide chain terminators in 1975 (*1*). Then the shotgun sequencing strategy was developed in the early 1980s (*2 – 4*). Shotgun sequencing has been the fundamental method for large-scale DNA sequencing in the last 20 years (*5 – 7*). However whole-genome shotgun sequencing (WGA) had been routine only in small organisms such as bacterial genomes. Lots of high copy repeats in the genomes of higher vertebrates make WGA at an enlarged risk of mis-assembly. In the initial stage of the human genome project, there were scientific debates over whether to use hierarchical shotgun sequencing (*8*) or WGA (*9*) to sequence the human genome. The International Human Genome Sequencing Consortium chose the latter (*10*). In recent years, there are some successful examples of WGA of complex eukaryotic genomes, such as that of *Drosophila melangoster* (*11*) and *Homo sapiens* (*12*) assembled by Celera Genomics, and that of *Oryza sativa* assembed by Beijing Genomics Institute (*13*, *14*). With extensive applications of capillary sequencers, the clone-end pair information is likely to reduce assembly trouble due to repeats. Therefore WGA will become the major method in genome research in the following years. Compared to "regional chromosome assembly", WGA can assemble random shotgun data without any high-density genetic or physical maps, so it has advantages of high speed, easy pipelining and little laborious headwork.

Here we present a statistical model, which inferred standards for identifying mathematically defined repeats (MDRs). With MDRs masked before assembly, the risk of mis-assembly was reduced and the speed of assembly was increased.

In the strategy of repeat-masked assembly, detected repeats were masked prior to the assembly, thus leaving so many gaps in the genome. For the purpose of gap closure, clone-inserts should cover re-

* **Corresponding author.**
E-mail: lisg@genomics.org.cn

peats at least twice, so it is necessary to design the length distribution of clone-inserts. First, clone-insert sizes affect repeat assembly. With proper clone-insert sizes, gaps were closed and the assembly was accurate (Fig. 1, A). If clone-inserts were not long enough to cover repeats, there would be many hubs linking to similar but distinct repeats. Thus we could not determine the correct path of assembly (Fig. 1, B). On the other hand, because of the varience(usually not less than $\pm 10\%$) of clone-inserts, if the length of clone-insert was more than 10 times that of repeats, we would mis-evaluate the accurate number of short tandem repeats (Fig. 1, C). Second, clone-insert sizes also affect scaffold construction. If clone-inserts were too short (approaching repeat length), it would lead to insufficient coverage and failure in bridging across contigs, gaps could not be closed (Fig. 1, D). On the contrary, too large clone-inserts would result in interleaving scaffold problems (Fig. 1, E), and we might incorrectly estimate the length distribution of gaps and of the total length of the genome.

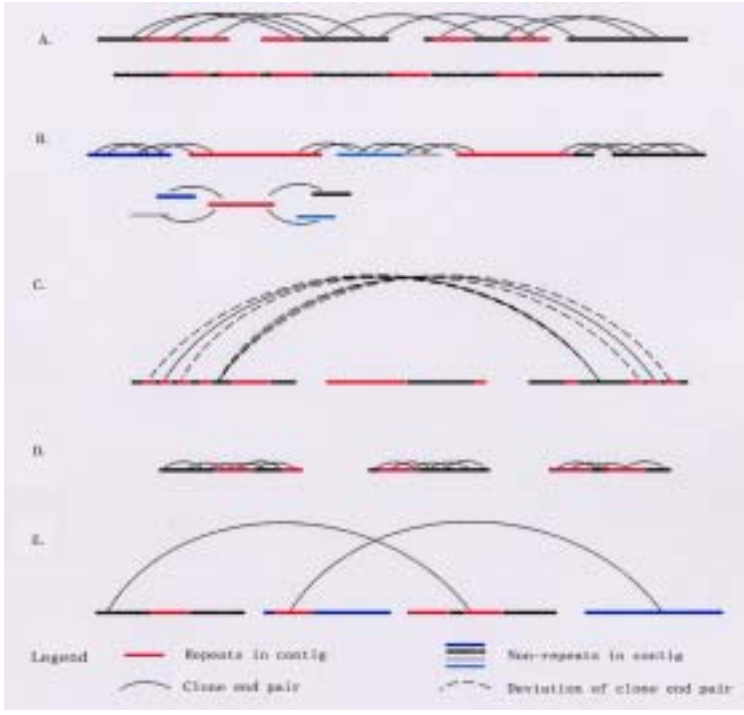

Fig. 1. The effect of clone-insert size on the accuracy of repeat assembly and scaffold construction. A. Repeats are assembled correctly with proper clone-insert size; B. If the clone insert is too short, the hubs linking to similar but distinct repeats make us unable to judge the correct path of assembly; C. If the clone insert is too long, we were not able to estimate the exact number of short tandem repeats because of the variation of clone-inserts, which is usually not less than $\pm 10\%$; D. Clone inserts being too short result in insufficient coverage, thus leaving too many gaps; E. Clone inserts being too large lead to interleaving scaffold problem.

Results

Recognizing MDRs using the model

From the model, we could deduce standards for recognizing MDRs at different shotgun coverage. Because 1-copy sequences (non-repeats) in shotgun data usually cover more than 2/3 of the genome, the probability of non-repeats being incorrectly defined as repeats should be controlled to a small (e.g. about 0.3%) number to avoid the false positive. Following this rule, we calculated a series of threshold depth by P_k (see Methods) and selected these depth (Table 1) as standards for recognizing MDRs at different coverage. If a 20-mer appeared more frequently than the standard, we regard it as a repeat at that coverage, otherwise as a non-repeat. When we use these standards, we could further reckon the false negative by G_{mk} (see Methods). We compared MDR detection efficiency at different standards for shotgun coverage of 1X and 4X

respectively (Fig. 2). And with the series of standards for recognizing MDRs, we also showed MDR detection efficiency at different shotgun coverages (Fig. 3). It is shown when coverage approached 4X and above, repeats with more than 5 copies could be detected, but increasing coverage contributed little to the efficiency of repeat detection.

To verify the feasibility and repeat-detecting efficiency of our model, we developed the software *MDRmasker* to find MDRs in the genome. *MDRmasker* was once applied to find repeats in simulated 2X, 4X, 4X+2X data of human and bacterial artificial chromosome (BAC) sequences of rice, which were described and discussed elsewhere (*13*, *14*).

In this paper, we focus on repeat detection efficiency and genome assembly result of a 6X simulated data of the human genome. We randomly selected high quality reads from 87 human BAC (Table 2), which came from a region of 11.9 Mb (3p24.3 to 3p26.1) on human chromosome 3. In the overlapping region, reads were picked out from only one BAC, so that the coverage of all the segments was equal. We merged all the data and established a 6X test set of human sequences. We run *MDRmasker* to detect repeats from the simulated data set (Fig. 4). After masking detected repeats with poly 'N' prior to assembly, we assembled the remaining data longer than 10 bp by Phrap. The information of clone-end pairs was further used to recover long repeats in the contigs and construct scaffolds (multiple contigs organized in correct order and orientation). Thus the assembly was finished (Table 3). We assembled the 6X test data set of human sequences by a straightforward use of Phrap as control. Unfortunately, the program died after a running of more than 400 h, which was more than ten times that used for repeat-masked assembly.

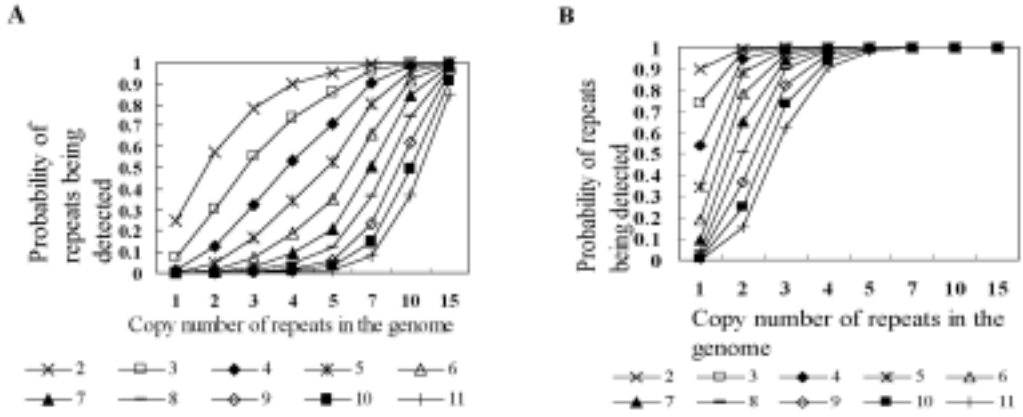

Fig. 2. The probability of detecting repeats of different copy number at different coverage and different depths. A. Results of a shotgun coverage of 1X; B. Results of a shotgun coverage of 4X. Each solid line depicts a depth, which may be a candidate standard.

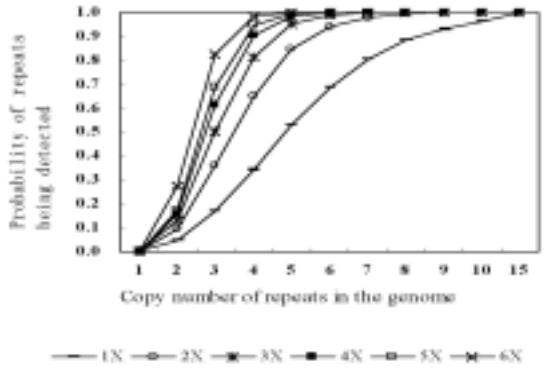

Fig. 3. The probability of detecting repeats as a function of their copy number in the genome at shotgun coverage of 1X, 2X, 3X, 4X, 5X, and 6X.

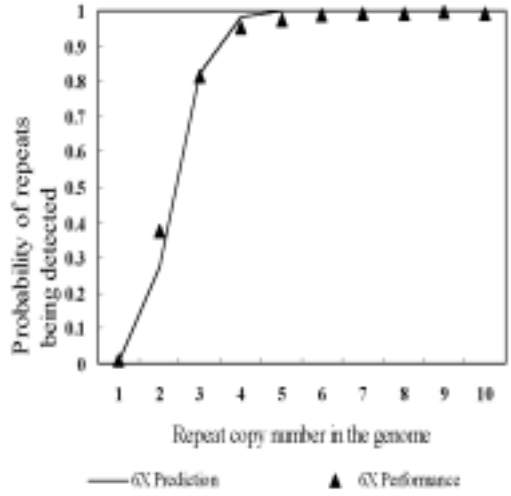

Fig. 4. The probability of detection of repeats in a 6X test set of human sequences. The solid line depicts the theoretical probability as predicted by the statistical model, and the triangles depict the actual probability detected by *ADRmarker*.

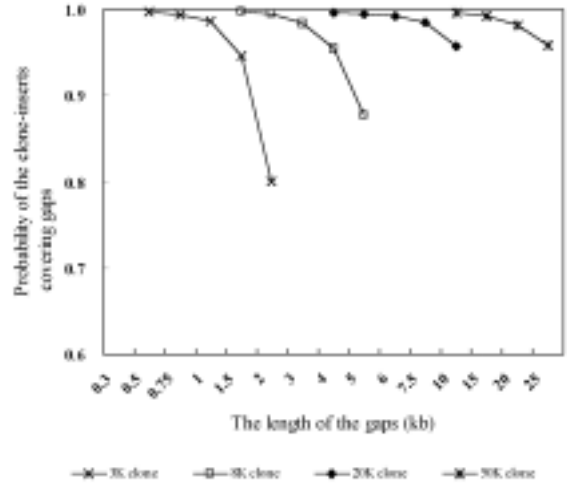

Fig. 5. With a 10X coverage of clone-inserts of 3 kb, 8 kb, 20 kb, and 50 kb, it shows the probability of clone-inserts covering gaps as a function of gap length at a 10X coverage.

Table 1 Standards for Identifying Repeats at Different Shotgun Coverage

Coverage	1X	2X	3X	4X	5X	6X
Repeat identifying standards	5	7	9	11	13	14

Table 2 Information of 6X Simulated Data Set of Human Sequences

Number of reactions	142,177
Genome size (Mb)	11.9
Average Q20 read length (bp)	528.88

Table 3 Assembly Results of 6X Test Set of Human Sequences

Number of contigs	2320
N50 contig length (kb)	10.42
Total genome size (kb)	12,165
Number of Singletons	5047
Coverage (X)	6.18
Misassembled Contigs	12
Frequency of mis-assembly	0.52%
Total length of mis-assembly (bp)	64,629
Length ratio of mis-assembly	0.53%

Table 4 The Recommended Length Distribution of Clone-Inserts without Considering the Length Distribution of Repeats

Clone-insert size (kb)	Coverage of clone-inserts (X)	Converted successfully Sequenced coverage(X)
50	10	0.2
20	10	0.5
8	10	1.25
3	10	3.33
0.5 *	1.3	1.3
Total		6.6

* Means either end of insert is successfully sequenced.

Table 5 Information of Simulated Genomes, for Which Length Distribution of Clone-Inserts Are Designed

	Human	Rice
Genome size (Mb)	11.9	4.5
Average Q20 read length (bp)	528.88	551.21

Table 6 Optimal Length Distribution of Clone-Inserts for Simulated Human Genome

Clone insert sizes (kb)	Coverage of clone inserts(X)	Converted successfully sequenced clones	Converted successfully sequenced reactions	Converted successfully sequenced coverage (X)
0.5*	1.1	26,999	26,999	1.2
3	9.0	35,700	71,400	3.2
8	8.3	12,346	24,692	1.1
20	10.0	5,950	11,900	0.5
50	0.0	2	4	0.0
Total			134,995	6

* Means either end of insert is successfully sequenced.

Table 7 Optimal Length Distribution of Clone-Insert for Simulated Rice Genome

Clone insert sizes (kb)	Coverage of clone inserts(X)	Converted successfully sequenced clones	Converted successfully sequenced reactions	Converted successfully sequenced coverage (X)
0.5*	1.2	11,526	11,526	1.201
3	8.7	13,920	27,840	2.9
8	9.5	5,700	11,400	1.188
20	9.5	2,280	4,560	0.475
50	12	1,152	2,304	0.24
Total			57630	6

* Means either end of insert is successfully sequenced.

Designing the length distribution of clone-inserts using the model

As a result of repeats masked, there were lots of gaps to be closed. Because the length of gaps is quite different, clone-inserts should be a series of sequences with different length, and the longest clone-insert should cover the longest repeat effectively. If the length distribution of repeats (gaps) in the target genome was not known, a coverage of 10X was recommended for each of the series of clone-inserts (Table 4), so that the probabilities of series of clone-inserts covering series of gaps at least twice could be all approaching 99% (Fig. 5). At this time, the accumulative successful sequencing coverage was about 6.6X.

For any given genome, the length distribution of repeats(gaps) could be calculated out. We made simulated genomes of human and rice (Table 5) and designed length distribution of clone-inserts for them. 53 BAC sequences of *Oryza sativa* L. ssp. *indica* were retrieved from http://www.tigr.org/tigr-scripts/IRGSP/Rstatus.cgi?chr=4&spp=indica, and concatenated as simulated rice genome. And the reference sequence of the 87 human BAC mentioned above was concatenated and used as simulated human genome. With length distribution of gaps in the simulated genomes reckoned by random sample method, we calculated the probability of series of clone-inserts covering gaps of different length using equations (*5*) and (*6*). Meanwhile, with the reckoned length distribution of gaps, we also estimated the expectation of uncovered gaps. Thus we knew how many base pairs were not covered effectively in a unit length of the genome. With the restriction of a given coverage (e.g. 6X), we could get an optimal length distribution of clone-inserts by a non-linear programming algorithm. In essence, the optimal length distribution of clone-inserts minimized the length of uncovered gaps. However, it was not necessary to keep the length distribution of clone-inserts at such a high precision obtained by non-lineal programming. So we manually adjusted length distributions of clone-inserts by Microsoft excel. The results of optimal length distribution of clone-inserts for man and rice are shown in Table 6 and Table 7, respectively. Repeats in simulated rice genome are much larger than those in simulated human genome. As a result, the optimal length distributions of clone-inserts for rice and human sequences were not the same. We set a minimal size of repeats so that repeats shorter than 200 bp were not considered. In the simulated human genome, repeats range from 0.2 to 6 kb; while in the simulated rice genome, they range from 0.2 to 25 kb. Therefore, clone-inserts were mainly 3, 8, and 20 kb for simulated human, and with an additional 50 kb for simulated rice.

Discussion

There were two reasons for choosing human sequences as simulated data to test *MDRmasker*. First, our genome is the most repetitive one we have at present and, second, we can further use the human finishing-map to test the validity of the repeat-masked assembly.

We could not assemble 6X test data of human sequences using common Phrap, because there were so

many repeats that Phrap failed to decide the correct assembly path from similar but distinct repeats. However, we successfully assembled the same data using the repeat-masked assembly strategy. After masking the repeats , the process of assembly was largely simplified and Phrap was able to handle much larger data sets than in a common assembly process. So the statistical model for recognizing repeats was of great importance in WGA of complex eukaryotic genomes.

The repeat-masked assembly process had a little discrepancy if the size of the data set was different. If the target genome was in the order of millions or tens of millions of base pairs (e.g. the 6X test set of human sequences is 12 Mb in size), the repeat-masked sequences were assembled straightforwardly. But if the target genome was more than hundred megabases (e.g. the rice genome is 466 Mb), we must cluster repeat-masked data into several groups by their homology and preliminarily assemble them within the groups. After that, we recovered repeats in the contigs according to the original data, and clone-end pairing information was used to re-assemble data between groups. Because the clustering of data might be improper, pairwise alignment of contigs by BLAST (*15*, *16*) were made to be de-redundant. Clone end pairing information was further used to construct the scaffold.

Methods

Herein we define some words used in our paper. Copy number means the times a sequence occurs in the genome. Coverage means the times a genome is represented in the shotgun data. Depth refers to the number of times a fragment appears in the shotgun data. Q20 read length refers to the length of high quality reads, the error probability of which is less than 10^{-2}.

We define G, L, N, F by

G = total genome length;

N = qualified reaction number;

L = average Q20 read length;

F = the minimal recognizable fragment length.

To simplify the model, two assumptions were made. First, we supposed each Q20 read length being equal to L, and thus average Q20 read length was also L. Second, repeats shorter than the minimal recognizable fragment length F, which was 15-20 bp in our model, would not be considered.

Depth of a single base in shotgun data

First, we defined a random variable Y_{ik} to describe the depth of a single base in shotgun data.

$$Y_{ik} = \begin{cases} 1, & \text{when Point } i \text{ has depth of } k \\ 0, & \text{otherwise} \end{cases}$$

Given a point $[i]$ with depth k, there would be k reads having a starting point in the region of $[i-L+1, i]$ in the genome; but the other $N-k$ reads would not have a starting point in that region. And the length of the region was L. If the starting points of all the reads were distributed randomly in the genome, the probability of the random variant Y_{ik} being equal to 1 was:

$$P(Y_{ik}=1) = C_N^k \Big(\frac{L}{G}\Big)^k \Big(1-\frac{L}{G}\Big)^{N-k} \tag{1}$$

Because the total genome length G was much larger than the average Q20 read length L, we did not consider the deviation of probability of both start and end $L-1$ base pairs in the genome. Thus equation (1) is tenable for each point in the genome and the mean of points of depth k is:

$$E(Y_k) = E\Big(\sum_{i=1}^{G} Y_{ik}\Big) = G \cdot C_N^k \Big(\frac{L}{G}\Big)^k \Big(1-\frac{L}{G}\Big)^{N-k}. \tag{2}$$

Especially, the mean of points of depth 1 is:

$$E(Y_1) = NL\Big(1 - \frac{L}{G}\Big)^{N-1}. \tag{3}$$

Depth of fragments of a given length in shotgun data

Second, we reckoned the depth of fragments of a given length in shotgun data. Due to lack of position information and relationship with the other fragments, it was not likely that we would recognize MDRs only from single base, so we considered fragments of a given length. Similar to the case of single bases, given a fragment (length F and starting point $[i]$) with depth k, there would be k reads having starting points in the region of $[i - L + F, i]$ in the genome, but the other $N - k$ reads would not have starting points in that region. So the equation changed into:

$$P(Y_{ik} = 1) = C_N^k\Big(\frac{L - F + 1}{G}\Big)^k\Big(1 - \frac{L - F + 1}{G}\Big)^{N-k}. \tag{4}$$

When considering fragments instead of single bases, we just needed to substitute $L - F + 1$ in equation (4) for L in equation (1). The corresponding conclusions remained unchanged.

In genome surveys, to avoid identical fragments occurring by chance, the choice of fragment length depends on the total genome size. For example rice has a total genome size of 430 Mb and a total number of segments of about 10^8, so we chose 20 bp as its fragment length. Thus there could be 4^{20} (about 10^{12}) kinds of unique 20-nucleotide oligomers (20-mers). As a result, identical 20-mers by chance would not occur. But if we consider a bacterial genome, with a total number of segments of about 10^6, 15 bp was long enough to avoid identical fragments by chance.

Depth of fragments of a given length in MDRs

Now we began to find the depth of fragments with a given length F in MDRs.

As inferred in 2.1, the mean of points with depth k is $E(Y_k)$in shotgun data. Here we defined P_k by

$$P_k = E(Y_k)/G,$$

where the variant P_k is the probability of points of depth k appearing in shotgun data.

As inferred in 2.2 the probability of fragments with a given length F has depth k in shotgun data, as long as we change the value L into $L - F + 1$. However, the probability mentioned above was actually the probability of non-repeats, since we had supposed that each fragment only appeared once in the genome. In fact, repeats of m copy number actually appeared at m different positions in the genome, the observed depth of which was the sum of depth at all of the m positions. For example, depth 0 means that each of the m positions has depth 0; and depth 1 means that at one position there is depth 1 and at the other positions depth 0; similarly, depth 2 means there were either depth 1 at two positions and depth 0 at the others, or depth 2 at one position and depth 0 at the others, etc. The probability of m-copy repeats having depth k in shotgun data was defined as G_{mk} :

$$\begin{aligned}
&G_{m0} = P_0^m \\
&G_{m1} = C_m^1 \cdot P_1 \cdot P_0^{m-1} \\
&G_{m2} = C_m^2 \cdot P_1^2 \cdot P_0^{m-2} + C_m^1 \cdot P_2 \cdot P_0^{m-1} \\
&G_{m3} = C_m^3 \cdot P_1^3 \cdot P_0^{m-3} + C_m^2 \cdot C_2^1 \cdot P_1 \cdot P_2 \cdot P_0^{m-2} + C_m^1 \cdot P_3 \cdot P_0^{m-1} \\
&\cdots\cdots\cdots \\
&G_{mj+} = 1 - G_{m0} - G_{m1} \cdots - G_{mj-1},
\end{aligned}$$

where G_{mj+} refers to the probability of m-copy repeats having depth j and above.

Designing length distribution of clone-inserts

Because repeat assembly and scaffold construction were both affected by clone-insert sizes, we tried to design the length distribution of clone-inserts using our statistical model. The successful sequencing ratio was usually 90%, even if each of the clone-inserts was sequenced from both ends. 20% of the clones would still have only one end sequenced correctly. Hence in our model, we prescribed that about 20% reads had sequence information of either end.

To construct a scaffold over repeats correctly, there should be more than 50 bp non-repeat sequence unmasked at both ends of the clone-inserts. Equation (4) was therefore revised to:

$$P_k = P(Y_{ik} = 1) = C_N^k \Big(\frac{L - F - 100}{G}\Big)^k \Big(1 - \frac{L - F - 100}{G}\Big)^{N-k}, \quad (5)$$

where P_k denotes the probability of clone-inserts with length L covering repeats (of start point $[i]$ and length F) k times; N refers to the total number of clone-inserts; and G is the length of genome.

Equation (5) was used to design the length distribution of clone-inserts. Not considering the initial and terminal regions of the genome, equation (5) was tenable for any point in the genome, so the subscript i could be ignored. For repeats to be covered at least twice, the probability was:

$$P_{2+} = 1 - P_0 - P_1. \quad (6)$$

Acknowledgements

We are grateful to our colleagues who made great efforts to sequence the human chromosome 3 and establish the rice genome working-draft. This work is supported by Beijing Genomics Institute.

References

1. Sanger, F. and Coulson, A. R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* 94: 441-448.
2. Anderson, S. 1981. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res.* 9: 3015-3027.
3. Gardner, R.C., *et al.* 1981. The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. *Nucleic Acids Res.* 9: 2871-2888.
4. Deininger, P. L. 1983. Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. *Anal. Biochem.* 129: 216-223.
5. Koop, B. F., *et al.* 1992. Organization, structure, and function of 95 kb of DNA spanning the murine T-cell receptor C alpha/C delta region. *Genomics* 13: 1209-1230.
6. Chissoe, S. L., *et al.* 1995. Sequence and analysis of the human ABL gene, the BCR gene, and regions involved in the Philadelphia chromosomal translocation. *Genomics* 27: 67-82.
7. Rowen, L., *et al.* 1996. The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. *Science* 272: 1755-1762.
8. Green, P. 1997. Against a whole-genome shotgun, *Genome Res.* 7: 410-417.
9. Weber, J. L. and Myers, E. W. 1997. Human whole-genome shotgun sequencing. *Genome Res.* 7: 401-409.
10. Lander, E. S., *et al.* 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.
11. Myers, E. W., *et al.* 2000. A whole-genome assembly of *Drosophila*. *Science* 287: 2196-2204.
12. Venter, J. C., *et al.*. 2001. The sequence of the human genome. *Science* 291: 1304-1351.
13. Yu, J., *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296: 79-92.
14. Wang, J., *et al.* 2002. RePS: a sequence assembler that masks exact repeats identified from the shotgun data. *Genome Res.* 12: 824-831.
15. Altschul, S. F., *et al.* 1990. Basic local alignment search tool, *J. Mol. Biol.* 215: 403-410.
16. Altschul, S. F., *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

Received: 13 January, 2003
Accepted: 20 January, 2003