BIOINFORMATICS

Vol. 18 no. 3 2002
Pages 379–388

Separation of nearly identical repeats in shotgun assemblies using defined nucleotide positions, DNPs

Martti T. Tammi[1], Erik Arner[1], Tom Britton[2] and Björn Andersson[1,*]

[1]Department of Genetics and Pathology, Rudbeck Laboratory and [2]Department of Mathematics, Uppsala University, Uppsala, Sweden

Received on June 22, 2001; revised on October 11, 2001; accepted on October 18, 2001

ABSTRACT

An increasingly important problem in genome sequencing is the failure of the commonly used shotgun assembly programs to correctly assemble repetitive sequences. The assembly of non-repetitive regions or regions containing repeats considerably shorter than the average read length is in practice easy to solve, while longer repeats have been a difficult problem. We here present a statistical method to separate arbitrarily long, almost identical repeats, which makes it possible to correctly assemble complex repetitive sequence regions. The differences between repeat units may be as low as 1% and the sequencing error may be up to ten times higher. The method is based on the realization that a comparison of only a part of all overlapping sequences at a time in a data set does not generate enough information for a conclusive analysis. Our method uses optimal multi-alignments consisting of all the overlaps of each read. This makes it possible to determine defined nucleotide positions, DNPs, which constitute the differences between the repeat units. Differences between repeats are distinguished from sequencing errors using statistical methods, where the probabilities of obtaining certain combinations of candidate DNPs are calculated using the information from the multi-alignments. The use of DNPs and combinations of DNPs will allow for optimal and rapid assemblies of repeated regions. This method can solve repeats that differ in only two positions in a read length, which is the theoretical limit for repeat separation. We predict that this method will be highly useful in shotgun sequencing in the future.

Contact: bjorn.andersson@genpat.uu.se

INTRODUCTION

The method of choice for genome sequencing has for several years been the shotgun approach, where random sub-clones are sequenced, followed by computer assembly of the sequence fragments in order to reconstruct the original sequence. The size and complexity of the shotgun projects that have been undertaken has increased drastically in recent years, due to whole genome shotgun sequencing of large genomes and to the repetitive nature of many genomes and genomic regions. A main source of problems is the inability of conventional shotgun assembly programs to correctly assemble nearly identical sequence repeats, longer than the length of the shotgun fragments. The sequences include both dispersed and tandem repeats. This problem is currently one of the main limitations of the shotgun method, and the need for a method to solve this issue has been stated and thus is currently an active area of research.

It is impossible to separate and assemble repeats if the unique sites are further apart than a read length. Thus, in order to separate almost identical repeats, the method used must be sensitive enough to detect most sequence differences, and it needs to determine that a sufficient number of differences are present in reads from repeat regions. Several attempts have been made to resolve this problem (e.g. Green, 1996; Kececioglu and Yu, 2001; Pevzner *et al.*, 2001). Of these methods, the first (Green) has been in use for several years, but is unable to separate nearly identical repeats without extremely high demands on sequence quality. The other two methods seem promising, but remain to be thoroughly tested.

We here present a novel method to correctly separate nearly identical repeats longer than a read length. The main strategy is to compute error probabilities for nucleotide positions that may represent a difference between repeat copies. To accomplish this we have to ensure that close to all available information in a shotgun data set is used. We do this by making several multi-alignments consisting of a read and all of its overlaps as defined by an assembly program.

In this way, the method can find differences between almost identical repeats and the information obtained can

[*]To whom correspondence should be addressed.

be used to improve shotgun assemblies by selectively using specific positions, defined nucleotide positions, DNPs, for the ordering of reads in the assembly. In this paper, we present the statistical calculations used to identify DNPs and to distinguish them from sequencing errors, and we propose a new method for shotgun assembly of repeated regions using DNPs.

Extensive testing of this method on simulated shotgun data is presented and the results clearly show that repeat regions containing 1% sequence differences between repeat copies can be separated, even when the maximum sequencing error is 11%. This method solves the repeat problem, since with the current methods for DNA sequencing this is close to the theoretical limit for repeat separation. We predict that this will make it possible to resolve most repetitive regions that are encountered in shotgun sequencing projects.

METHODS

To separate nearly identical repeats, we need to detect unique differences between repeat units. To accomplish this, we construct multi-alignments consisting of sequence reads and all of their overlaps with other reads. In this way, we ensure that we always use as much information as possible when analyzing a repeat region. The differences between repeat units can be distinguished from sequencing errors by the fact that the errors are distributed randomly, whereas the real differences are not. However, the distributions computed on one column will overlap to varying degrees depending on the rate of sequencing error, coverage, the number of repeats, and the number of differences between repeat units. The separation of these distributions is not sufficient for detection of an acceptable rate of true positives. A better separation can be accomplished when we also consider the rate of coinciding deviations from column consensus between at least a pair of columns in the multi-alignments. This means that at least two differences need to be present in a read in order to be detected with our method.

The requirement of two differences present on a sequence read is a universal constraint on how similar the repeats can be in order to be separable. This is independent of the method used to assemble almost identical repeats longer than the read length and hence not a constraint imposed by our method. In order to elongate a contig, a read must contain at least two differences as illustrated in Figure 1. It is mandatory to detect almost all the differences that are present in the template sequence, since by definition almost identical repeats contain only a few differences. Thus, to separate repeats it is necessary to use a method that with high confidence can establish whether at least two differences are present in a read. The following sections describe the method in detail.

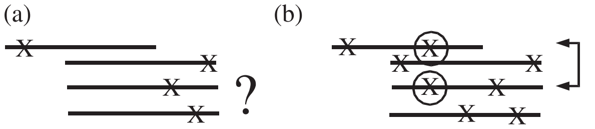

Fig. 1. When assembling fragments from nearly identical repeats, the detected differences must be used to determine which reads belong together. (a) No difference is present along the alignment and it is hence impossible to determine which reads should be joined. (b) The first and the third read share a difference and can be joined. Bars indicate reads, an X indicates a detected difference.

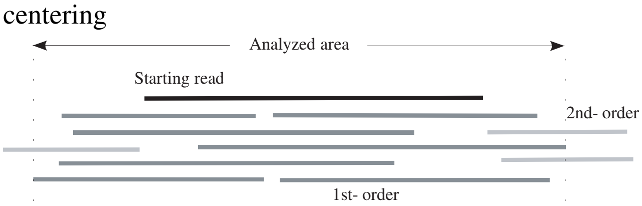

Fig. 2. Schematic view of analyzed overlaps in a multi-alignment. Dark grey bars indicate 1st order overlaps, light grey bars indicate 2nd order overlaps, and the black bar indicates the starting read. The analysis is performed on the starting read and the reads with 1st order overlaps.

Preparation of multi-alignments

To analyze a repeat region, a starting read from the data set is chosen and a multi-alignment is constructed from the read, all its overlaps with other reads, 1st-order overlaps, and all their overlaps with reads not already included in the alignment, 2nd-order overlaps (Figure 2). The alignment is optimized locally using the ReAligner algorithm (Anson and Myers, 1997). After analysis, the starting read and the reads with 1st-order overlaps are marked as analyzed, and a new starting read is picked from the set of non-analyzed reads. This process is repeated until all reads in the data set have been marked as analyzed.

Detection and pair wise evaluation of candidate columns

The consensus base in a column is defined as the most frequent base in the column. The candidate columns are those where deviations from the consensus are observed and where these deviations contain at least $D_{\min}$ bases of the same type, a candidate base type. Two different methods have been tested: the basic and the extended. The goal of the basic method is to locate pairs of candidate columns where the deviations coincide in at least $D_{\min}$ reads covering both columns. In the extended method, a computation of the probability of observing the coinciding deviations by chance is added.

Consider two fixed positions, u and v, in a multi-alignment of k sequences spanning both u and v. Denote the base at position u for the jth sequence by $a_{u,j}$ and similarly $a_{v,j}$ for position v. Let $I_{u,j}$ be the indicator for the event that the base at position u of the read j deviates from consensus and is a specific candidate base type. That is, $I_{u,j} = 1$ if this is the case and $I_{u,j} = 0$ otherwise. The total number of deviations from consensus at position u is denoted $N_u = \sum_{j=1}^{k} I_{u,j}$. Define $I_{v,j}$ and N_v in a similar way. Let $I_j = I_{u,j} I_{v,j}$ be the indicator for a 'coincidence' of the jth sequence, i.e. that both positions deviate from consensus in a sequence j. Finally, the total number of coincidences is denoted as:

$$C = \sum_{j=1}^{k} I_j = \sum_{j=1}^{k} I_{u,j} I_{v,j}.$$

In the present section, we derive a test for the hypothesis that coincidences in the sequences occur by chance, compared to the alternative that they appear systematically. The systematic appearance of coincidences indicates that certain positions contain information about differences between repeat copies rather than sequencing errors.

Below we compute the approximate distribution of C under the assumption that all deviations from the consensus occur independently. The probabilities $p_{u,j} = P(I_{u,j} = 1)$ and $p_{v,j} = P(I_{v,j} = 1)$ are computed from Phred quality values (Ewing and Green, 1998).

In order to test if the number of coincidences C is large due to chance and not because there were more deviations from consensus (at the two positions) than expected, we derive the distribution of C given the observed values $N_u = n_u$ and $N_v = n_v$, i.e. the number of deviations from the consensus at the two positions.

To derive the exact distribution of C conditional on $N_u = n_u$ and $N_v = n_v$ in a typical shotgun data set is very complicated except for the case when $p_{u,1} = \cdots = p_{u,k}$ and $p_{v,1} = \cdots = p_{v,k}$. When the ps are identical in all sequences, standard combinatorics imply that the distribution of C given n_u and n_v is hypergeometrically distributed with parameters k, n_u, and n_v/k. This means that

$$P(C = x) = \frac{\binom{n_v}{x}\binom{k - n_v}{n_{u-x}}}{\binom{k}{n_u}}, 0 \leqslant x \leqslant n_v,$$
$$0 \leqslant n_u - x \leqslant k - n_v.$$

This is true, since when all ps are identical each possible configuration has equal probability. By considering the n_v deviations from the consensus in position v as fixed, the denominator above gives the total number of ways to distribute n_u deviations among all k sites, and the numerator the number of ways to do this resulting in x coincidences. When the different p values are not identical, it is very cumbersome to compute the distribution of C. The reason for this is that it is necessary to divide the conditioning event $N_u = n_u$ (or $N_v = n_v$) in the different ways this event can occur, and since each such event has a separate expectation of C and a separate probability.

While the case where all p values at both sites are identical is straightforward, all other cases require an approximation of the distribution. Because all the p values are assumed small (typically of the order 0.1 or smaller) the unconditional distribution of C is well approximated by the Poisson distribution (Ross, 1988). On the other hand, the conditioning on N_u and N_v only introduces weak dependencies, which implies that a Poisson approximation should still be satisfactory. What remains is to compute the mean parameter of the Poisson distribution, i.e. compute $E(C|N_u = n_u, N_v = n_v)$. From the definition of C follows

$$\begin{aligned} &E(C|N_u = n_u, N_v = n_v) \\ &\quad = \sum_{j=1}^{k} E(I_{u,j}|N_u = n_u) E(I_{v,j}|N_v = n_v) \\ &\quad = \sum_{j=1}^{k} E(I_{u,j} = 1|N_u = n_u) E(I_{v,j} = 1|N_v = n_v). \end{aligned}$$

Further,

$$\begin{aligned} P(I_{u,j} = 1|N_u = n_u) &= \frac{P(I_{u,j} = 1, N_u = n_u)}{P(N_u = n_u)} \\ &= \frac{P(I_{u,j} = 1, N_u^{(j)} = n_u - 1)}{P(I_{u,j} = 1, N_u^{(j)} = n_u - 1) + P(I_{u,j} = 0, N_n^{(j)} = n_u)}, \end{aligned}$$

where $N_u^{(j)} = N_u - I_{u,j}$ denotes the total number of deviations from the consensus at site u excluding read j. Note that $I_{u,j}$ and $N_u^{(j)}$ are independent. Furthermore, both N_u and $N_u^{(j)}$ are approximately Poisson distributed. Let $\lambda_u = \sum_{i=1} p_{u,i}$and $\lambda_u^{(j)} = \lambda_u - p_{u,j}$, respectively, denote the means for these Poisson distributions. It follows that:

$$\begin{aligned} \frac{P(I_{u,j} = 1, N_u^{(j)} = n_u - 1)}{P(N_u = n_u)} &\approx \left(p_{u,j} \mathrm{e}^{-\lambda_u^{(j)}} \lambda_u^{(j)^{n_u - 1}} / (n_u - 1)!\right) \\ &\Big/ \left(p_{u,j} \mathrm{e}^{-\lambda_u^{(j)}} \lambda_u^{(j)^{n_u - 1}} / (n_u - 1)! + (1 - p_{u,j}) \mathrm{e}^{-\lambda_u^{(j)}} \lambda_u^{(j)^{n_u}} / n_u!\right). \end{aligned}$$

This becomes:

$$\frac{P(I_{u,j} = 1, N_u^{(j)} = n_u - 1)}{P(N_u = n_u)} \approx \frac{n_u p_{u,j}}{n_u p_{u,j} + \lambda_u^{(j)}(1 - p_{u,j})}.$$

The corresponding result applies to $P(I_{v,j} = 1|N_v = n_v)$, and we conclude that

$$E(C|N_u = n_u, N_v = n_v) \approx \sum_{j=1}^{k} \frac{n_u p_{u,j}}{n_u p_{u,j} + \lambda_u^{(j)}(1 - p_{u,j})} \times \frac{n_v p_{v,j}}{n_v p_{u,j} + \lambda_u^{(j)}(1 - p_{v,j})}. \quad (1)$$

Hence, the suggested approximation of the distribution of C given $N_u = n_u$ and $N_v = n_v$ is to approximate it with the Poisson distribution having the mean specified above. This implies that the hypothesis that coincidences occur by chance rather than for systematic reasons can be tested by comparing the observed value of c_{obs} with what to expect from the derived approximate distribution. We compute

$$p^{\text{corr}} = 1 - \sum_{i=0}^{c_{\text{obs}}-1} Po(i),$$

where $Po(i)$ is the probability function for a Poisson variable with mean $E(C|N_u = n_u, N_v = n_v)$. p^{corr} is the probability of observing c_{obs} or more coincidences between columns u and v. The hypothesis is accepted if p^{corr} exceeds $p^{\text{corr}}_{\text{max}}$.

To minimize the effect of columns with a large number of expected sequencing errors, we compute the probability of observing any of the two columns by chance, p^{col}:

$$p^{\text{col}} = p_u^{\text{col}} + p_v^{\text{col}} - p_u^{\text{col}} p_v^{\text{col}}$$

where $p_u = 1 - \sum_{i=0}^{n_u-1} Po(i)$, and $Po(i)$ is the probability function for a Poisson variable with mean $\varepsilon_u = \sum_{i=1}^{k} p_{u,i}$. p_v is computed in a similar fashion. We compute $p^{\text{tot}} = p^{\text{col}} p^{\text{corr}}$ and reject the pair u, v if p^{tot} exceeds $p^{\text{tot}}_{\text{max}}$. Otherwise, the positions are assigned as DNPs, defined nucleotide positions. These positions represent differences between repeat units and can therefore be used in order to separate the repeats.

RESULTS

A method for separation of repeats in shotgun sequence data has been developed. In this strategy, shotgun reads are compared in multi-alignments, which allows for the identification of sequence differences between repeat units that can be used for a correct assembly. In order to show the efficiency of the method and to test its limits, we have used several simulated shotgun projects that were produced using the programs ‘gen_seq’ and ‘sim_gun’, that were developed in house for testing fragment assembly programs. Both programs are available from the authors. The gen_seq program produces random DNA sequences with a specified number and length of inserted repeats. A specified number of randomly placed differences between repeats can also be introduced. The output is read by sim_gun, a program that simulates a shotgun sequencing process.

In order to evaluate the method, five sets of simulations were performed. The simulations were designed to closely mimic real shotgun data, and for that reason, the error rates in each set were imported from nine different real shotgun projects as compared to simpler methods using flat error rates. For each such shotgun project quality file, we simulated projects consisting of repeat units of length 1000, 2000 and 3000 bases repeated 4, 6, 8 and 10 times in tandem, resulting in 108 assemblies per set. The first three sets of simulations were performed at different levels of quality trimming of the simulated sequence reads, which resulted in varied coverage and read length. The fourth set used the same quality trimming as the third set, but with higher coverage. In the final simulation, the number of reads was decreased, in order to reduce the coverage. The difference between any two repeats was 1.0% consisting of randomly distributed base substitutions. A control set was added, in which the repeats were identical. In total, 648 simulations were performed including the control set. The properties of the simulation sets are listed in Table 1.

Distributions of sequencing errors and real differences computed on one column only

The probability distributions of sequencing errors compared to those of true differences were examined. These distributions were found to overlap. As an example of this, Figure 3 shows one subset of distributions from Sim 3 containing ten repeats in tandem. At least six or seven differences from the consensus sequence must be observed in order to separate true differences from sequencing errors. We define a true difference in a read to be a position corresponding to a difference between repeat copies in the template, that is the target DNA molecule, as compared to a change caused by sequencing error.

The expected number of coincidences due to sequencing errors and real differences

To test the reliability of the approximation of the number of coincidences (1), we computed the expected number of coincidences over two columns and compared it to the observed number of coincidences. The means are well separated and the observed means of coincidences due to sequencing errors agree with the approximation, except for means that are greater than one. The reason for this is that there were few cases where the expected number of coincidences due to sequencing errors was greater than one (data not shown).

Table 1. The five different shotgun simulation sets and control set. The quality values used in simulations originate from nine different real sequencing projects

Simulation	Number of assemblies	Average sequencing error after trimming	Maximum sequencing error allowed in trimming (%)	Coverage after trimming	Average read length	Number of reads after trimming
Sim 1	108	4.3	11	8.7	494	26 491
Sim 2	108	3.3	8	7.6	472	23 838
Sim 3	108	2.6	6	6.8	459	22 423
Sim 4	108	2.6	6	10.2	457	33 866
Sim 5	108	2.6	6	3.5	463	11 323
Control	108	4.3	11	8.7	494	26 491

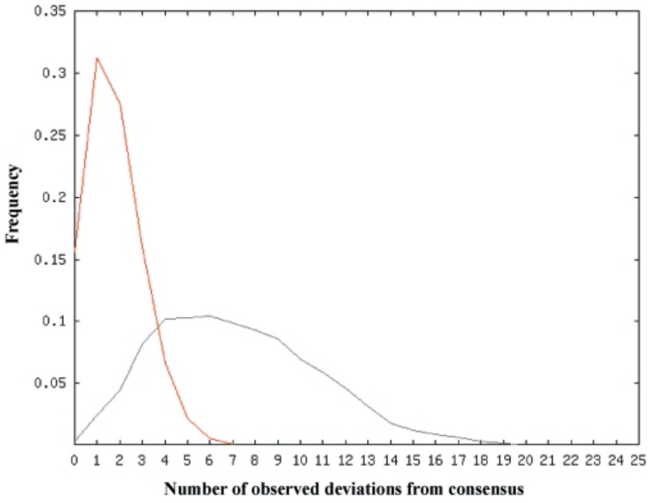

Fig. 3. The distribution of sequencing errors and real differences computed on one column using a subset of Sim 3 consisting of 1000, 2000, and 3000 bases long repeat units repeated ten times in tandem, computed on columns where the expected number of errors is one. The difference between any two repeat units is 1%. The distribution of sequencing errors is on the left side, and the distribution of real differences on the right.

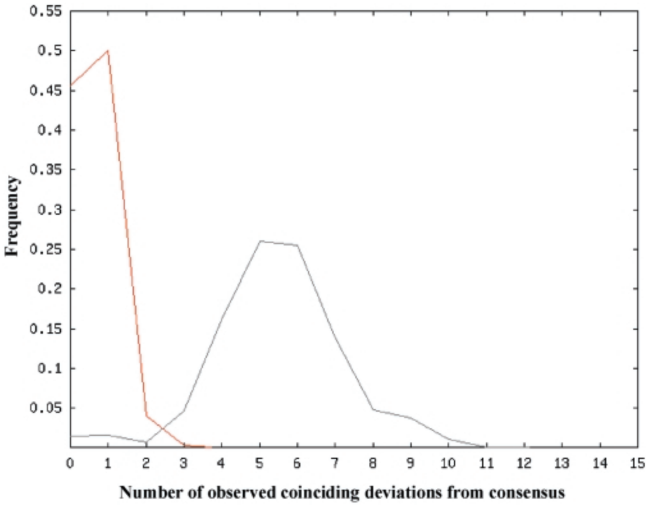

Fig. 4. The distribution of coincidences due to sequencing errors and real differences computed on two columns using a subset of Sim 3 consisting of 1000, 2000, and 3000 bases long repeat units repeated ten times in tandem, computed on pairs of columns where the expected number of coincidences is one. The difference between any two repeat units is 1%. The distribution of sequencing errors is on the left side, and the distribution of real differences on the right.

The distribution of coincidences of sequencing errors and real differences computed on two columns

The distributions of coincidences due to sequencing errors and due to differences between repeats were computed and examined. Figure 4 shows an example of such distributions. These two distributions overlap only up to three coincidences. Thus, in this example, four or more coincidences are enough in order to separate errors from differences.

Since several properties, for example the quality of the reads, the mean read length, coverage and the number of repeats, may influence the performance of the methods, we tested both the basic and extended methods on varied data sets.

Basic method

Table 2 shows the results of the basic method, where no probability values are calculated, applied to Sims 1–3. These simulations differ only by the stringency of the trimming, resulting in different read mean qualities, number of reads after trimming and coverage. The error percentage, i.e. the rate of false positives, and the sensitivities are shown for different $D_{\min}$. For comparison, two different sensitivities are computed for each $D_{\min}$: the sensitivity in respect to true differences in reads and to differences in the template sequence, S_R and S_T, respectively, where

$$S_R = 100 \times \frac{\text{number of detected differences}}{\text{total number of true differences}}.$$

Table 2. The results of the basic method in Sims 1–3. The error, ε, sensitivity in respect to true differences in reads, S_R, and sensitivity in respect to differences in the template, S_T, at different $D_{\min}$. S_R is set to 100 in Sim 1, $D_{\min} = 2$, for comparison

	$D_{\min}$				
	2	3	4	5	6
Simulation	(ε)	(ε)	(ε)	(ε)	(ε)
	S_R/S_T	S_R/S_T	S_R/S_T	S_R/S_T	S_R/S_T
	(%)	(%)	(%)	(%)	(%)
Sim 1	(59) 100/97	(4.3) 89/87	(0.55) 81/78	(0.42) 71/65	(0.36) 60/53
Sim 2	(40) 81/94	(2.6) 73/82	(0.37) 64/70	(0.28) 55/57	(0.26) 43/43
Sim 3	(25) 71/90	(0.71) 62/76	(0.27) 54/62	(0.21) 44/48	(0.20) 33/34

S_T is computed similarly. For comparison purposes the total percentage of true differences in reads is set to 100 for $D_{\min} = 2$ in Sim 1. The error decreases from Sims 1 to 3 and from $D_{\min} = 2$ to 6, as well as the total number of both true template and read differences detected. In general, a decrease in S_R corresponds to a decrease in S_T. Thus, with lower constraints on sequence quality, more true differences will be detected. If the quality clipping is more stringent, the error decreases, but a larger number of true differences in the template remain undetected.

The quality clipping in Sims 4 and 5 is identical to that used in Sim 3. The difference between these simulations is the number of reads simulated, resulting in higher and lower shotgun coverage, respectively. The results of the basic method for these data sets are shown in Table 3. The error and the sensitivity in respect to reads are shown for different $D_{\min}$. The error decreases in both Sims 4 and 5 with increasing $D_{\min}$. Compared to Sims 3 and 4, very few true differences are detected in Sim 5 due to the low coverage. In comparison, in Sim 4, the high coverage results in detection of a higher proportion of the true differences compared to Sim 3. It is thus clear that increased shotgun coverage, due to either less stringent quality clipping or an increased number of reads, results in improved detection of differences between repeat units.

Table 3. The results of the basic method in Sims 4–5. The error, ε, and sensitivity in respect to true differences in reads, S_R, at different $D_{\min}$

	$D_{\min}$				
	2	3	4	5	6
Simulation	ε/S_T	ε/S_T	ε/S_T	ε/S_T	ε/S_T
	(%)	(%)	(%)	(%)	(%)
Sim 4	30/95	1.3/89	0.21/83	0.17/78	0.14/69
Sim 5	18/64	0.69/41	0.47/23	0.47/21	0.77 / 11

Table 4. The error of the basic method on subsets of Sims 1–5 containing only four and ten repeat units

	$D_{\min}$				
Simulation	2	3	4	5	6
	Error (%) 4/10 repeats				
Sim 1	56/61	4.1/5.1	1.1/0.51	0.99/0.31	0.75/0.22
Sim 2	40/41	2.3/1.5	1.2/0.26	1.1/0.14	0.94/0.073
Sim 3	28/26	2.1/0.52	1.4/0.14	1.3/0.069	1.1/0.059

The effect of different repeat copy numbers on the results was tested on Sims 1–3 by extracting and comparing simulations with four and ten repeat copies. In most cases, the error was found to be lower in sets containing ten repeats compared to sets that contained four repeats (Table 4). One possible explanation for this is that with fewer repeat units, the risk of choosing the wrong consensus sequence in certain columns in the multi-alignments is increased. This is especially true in the case of two repeat copies, but this has not yet been investigated. The difference is probably not sufficient to necessitate the use of different models for different repeat copy numbers. A similar effect was observed using the extended method (data not shown).

Extended method

We applied the extended method with $p^{\text{tot}}_{\max} = 10^{-3}$ to Sims 1–3 (Table 5). The sensitivities in respect to reads, S_R, are shown. A comparison of the basic and extended methods at $D_{\min} = 2$, showed that the error decreased 85% in Sim 1 and 18% in Sim 3 when the extended method was used. At $D_{\min} = 3$ the error decreased 63 and 27%, respectively, while the loss in sensitivity in respect to reads varied from 9 to 0%. No significant reduction of the error rates was achieved by using the extended method when $D_{\min} > 3$. An examination of the errors remaining at $D_{\min} > 3$ suggested that the vast majority were caused by alignment errors, and that the rest consisted of bases erroneously 'sequenced' as the same base as a corresponding difference in another repeat copy, thus indistinguishable from a real difference.

The probability values for the analysis are important for the success of the extended method. Different $p^{\text{tot}}_{\max}$ yields different results regarding the error rate and the sensitivity. A lower $p^{\text{tot}}_{\max}$ resulted in a lower error rate and a lower sensitivity in all simulations (data not shown). Correspondingly, a higher $p^{\text{tot}}_{\max}$ resulted in a higher sensitivity and a higher error. Figure 5 shows the variation in sensitivity and error for different choices of $p^{\text{tot}}_{\max}$ in Sim 1 at $D_{\min} = 3$. Note that a choice of $p^{\text{tot}}_{\max} = 1$ is

Table 5. The results of the extended method in Sims 1–3. The error at $p_{\max}^{\text{tot}} = 10^{-3}$ ($\varepsilon_{10^{-3}}$), and corresponding sensitivity with respect to true differences in reads (S_R), computed at different $D_{\min}$. S_R is computed in relation to Sim 1, $D_{\min} = 2$ with the basic method

	$D_{\min}$				
	2	3	4	5	6
Simulation	$\varepsilon_{10^{-3}}$ (S_R) (%)	$\varepsilon_{10^{-3}}$ (S_R) (%)	$\varepsilon_{10^{-3}}$ (S_R) (%)	$\varepsilon_{10^{-3}}$ (S_R) (%)	$\varepsilon_{10^{-3}}$ (S_R) (%)
Sim 1	8.9 (91)	1.6 (87)	0.53 (81)	0.42 (71)	0.36 (60)
Sim 2	6.6 (76)	0.92 (72)	0.37 (64)	0.28 (55)	0.21 (43)
Sim 3	5.5 (67)	0.52 (62)	0.27 (54)	0.21 (0.44)	0.20 (0.33)

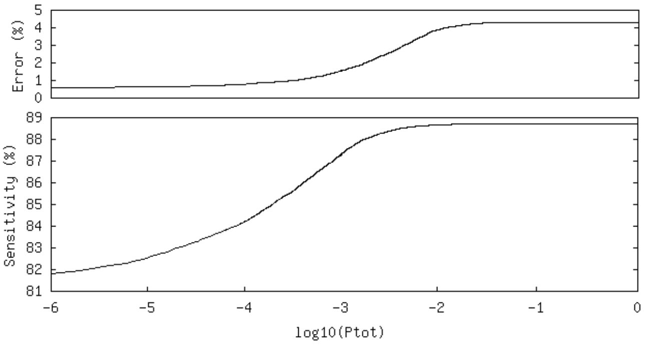

Fig. 5. The variation in error and sensitivity of the extended method at different $p_{\max}^{\text{tot}}$ in Sim 1, $D_{\min} = 3$.

identical to using the basic method. These results indicate that the extended method is flexible and can be readily optimized, and that it can be used to reduce the error of the basic method without a significant reduction of sensitivity.

In the control set, the basic method erroneously detected 58 298 DNPs at $D_{\min} = 2$. With $D_{\min} = 3$, only 693 or 1.2% of these erroneous DNPs remained. Using the extended method with $p_{\max}^{\text{tot}} = 10^{-3}$, 162 or 0.3% remained, and with $p_{\max}^{\text{tot}} = 10^{-5}$, no DNPs were erroneously detected in the control set. Thus, the use of the extended method significantly reduces the occurrence of false positives.

DISCUSSION

A main limitation of the shotgun sequencing strategy has been its inability to correctly resolve repetitive sequences. It is of course impossible to resolve completely identical repeats. However, in most cases base differences between repeat units are present. While no method that resolves nearly identical repeats has yet been published, the need for such a method has been stated and it has been suggested that one way to approach the problem is to use single base differences between repeat copies. We here present the first strategy capable of successfully resolving repeats. This method shows a significant improvement over existing methods according to the rigorous testing performed.

The results shown are based on probabilities computed on column pairs, which implies that at least two true differences must be present in a sequence read in order to be detectable by our method. However, regardless of the method used to detect differences, two differences are required to be present in a read in order to assemble it correctly (Figure 1). Thus, if the average read length is e.g. 500 base pairs, the lowest percentage of unique differences between repeat copies must theoretically be at least about 0.4% for these to be separable. In reality, however, more differences are required, since the probability is extremely low that several reads simultaneously cover two positions as far apart as the mean read length. Furthermore, the differences will not be evenly distributed across the repeat region. Our intention has been to study the limits of the method. For this reason we have used coincidences between two columns. It is clear that in the analysis of reads differing more than the minimum amount, the computation can be based on more than two columns, which yields more accurate results.

This method uses multi-alignments and identifies variable sites using the calculation of probabilities of the occurrence of certain combinations of sequence differences. We have shown that the computation of probabilities on one column only, even if we use all the available overlaps, lacks the statistical power to accurately pinpoint the unique differences between repeat copies (Figure 3). For this reason, we perform the computation on several columns at the same time, which gives a good distinction between sequencing errors and differences between repeats (Figure 4). From Table 2 it is clear that it is possible to distinguish differences from sequencing errors with good accuracy using the basic method. It is also clear that the error decreases with more stringent quality clipping. The cost for this increase in accuracy is a decreased number of detected differences. As Table 2 shows, almost twice as many differences in the template remain undetected in Sim 3 compared to Sim 1 if we use the basic method with $D_{\min} = 3$ (24 versus 13%). Since an assembly may break at every undetected difference using the assembly method outlined below, an attempted assembly of fragments from repeats differing only 1% would potentially give rise to twice as many contigs under the trimming conditions in Sim 3 as for those in Sim 1. Using the same quality constraints as in Sim 3, the results from Sim 4 show that a 50% increase in the number of sequence reads is necessary to obtain a similar performance as in Sim 1 regarding detected template differences (Tables 1 and 3). The error in Sim 1 is albeit higher than in Sim 4 (4.3 versus 1.3% for $D_{\min} = 3$), but this error can be decreased to the level of Sim 4, without a substantial loss in detected differences, by applying the extended method as shown in Table 5.

From Table 5 it is apparent that the extended method has a greater effect when the mean read qualities are lower due to less stringent quality clipping. This is what would be expected, since a high constraint on read quality removes most of the sequencing errors, making the probability of observing multiple coincidences on two columns due to chance close to zero. In this case, the erroneously detected differences occur mainly because of alignment errors. Using the basic method, the problem of separating repeats is a trade-off between sensitivity and accuracy, i.e. the longer the mean read length after trimming, the more differences we can detect but we also increase the error rate since a longer read length leads to a lower mean quality. The extended method is a means to address this problem and to allow for a lower error without a substantial decrease in sensitivity.

Clustering reads using DNPs

The ultimate goal is to use this method in an assembly program, and we have previously developed a preliminary version of an assembly program using DNPs (Tammi *et al.*, 2001). There are several possible ways to use the results of the method to direct shotgun assembly. We outline one such approach. The reads in a multi-alignment can be clustered into distinct repeat groups using the DNPs in individual reads. This will not add significantly to the runtime, since the multi-alignment can be produced in linear time from the pair wise alignments. The most computationally expensive step is the optimization of the alignment. Optimally, the repeats are separated as shown in Figure 6a. In this case, the problem of assigning reads to different repeat groups is relatively straightforward. However, in some cases, false positives may create conflicts between groups. One such example of a conflict is shown in Figure 6b, where one read can be assigned to two repeat groups. This can be handled in different ways, for example: (1) to do an exhaustive computation of which group compositions are the most probable; (2) to simply discard the conflicting reads. This can also be motivated if the objective is to make the best possible assembly with as few errors as possible.

In cases, when repeats are not identical, a distinction between fragments originating from unique and repeated regions can be made using DNPs. Once the reads are clustered into separate groups, contig construction can be performed by chaining clusters together. In this process, three requirements are needed to chain two clusters: (1) the consensus of the clusters must match to a certain degree along the alignment. This requires no additional matching, since the analysis stage defines which reads, and thereby which clusters, that match; (2) the clusters must match at a DNP; (3) if no DNP is present along the alignment of two clusters, they may not be joined. This allows for a rapid assembly once the previous analysis has been performed, and an indication of where more sequencing is needed, since gaps will form in regions where no DNPs have been detected due to low coverage.

It is important that the DNPs are unique when we use them to chain the clusters to assure correct order. Non-unique DNPs may occur, that is, when the same difference occurs in more than one repeat unit. When accompanied by unique DNPs, non-unique DNPs can be detected as shown in Figure 6c. In ambiguous cases, we can compute a probability that the observed coverage in the region comes from only one repeat copy using the binomial distribution, and assign the DNP as non-unique if the probability exceeds a threshold.

Considerations for repeat separation

Separating nearly identical repeats is a task that imposes high demands on precision. When shotgun fragments contain as few as two differences along a read length, it is necessary to detect them and use them in contig building. A method that scores overlaps based on sequence similarity and high quality mismatches, but does not explicitly identify the differences, will ultimately fail if the repeats are too similar.

Several algorithms have been developed in order to separate repeated sequences in shotgun assembly. The existing algorithms can be successful in separating short repeats, and even repeats longer than the mean read length, when the amount of differences present is sufficiently high, in practice more than 2–3%. When the repeats are more similar, these methods fail. Phrap (Green, 1996), for example, uses error rates and statistical models to separate repeats. The overlaps are sorted and additional comparisons are performed at the contig layout stage. This method, while successful when read qualities are high and the repeats differ to a high extent, lacks statistical power when only few differences are present in the data set, since it is based on read pairs and thus does not use all information available. Further, it does not attempt to evaluate the number of real differences present in reads and pinpoint them. This most often results in assemblies, where sequence fragments originating from different repeat copies are erroneously joined together.

A recently described algorithm (Kececioglu and Yu, 2001) is based on locating single base differences to separate repeats. The approach is to perform the analysis of single base substitutions on contigs produced by a fragment assembly program. The analysis is performed

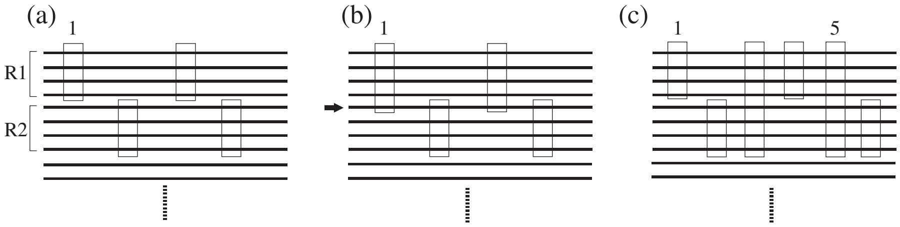

Fig. 6. A schematic view of multi-alignments containing DNPs. (a) Optimally separated repeat copies R1, R2. (b) A conflicting read. (c) Optimally separated repeat copies containing non-unique DNPs. The boxes indicate DNPs.

iteratively. Using this method, important information may be lost for the reason that repeats may be partly resolved before the analysis is performed, which makes the analysis of single base substitutions between fragments incomplete. Another approach would be to use the algorithm on multi-alignments, as in our method, before the contig construction stage. Similar to our method, the algorithm looks for correlations at multiple columns to assess the status of bases differing from consensus. As opposed to our method, an upper bound on the rate of difference between repeat copies is needed, but the major difference is that this algorithm, like phrap, computes statistics on pairs of sequence reads, albeit in a multi-alignment. Our method, on the other hand, performs the analysis simultaneously on all reads covering two candidate positions. This method uses more than two columns where applicable, and while our method has not been tested using more than two columns, when the repeats differ as little as 1% between any two repeat copies, there will in most cases be only two differences present in a read. Thus, we can directly compare this method to our method in the case when we encounter a pair of columns where two correlating differences from consensus are observed. As we have shown above, accepting as few as two correlating differences as DNPs yields an unacceptably high amount of false positives. The method described in Kececioglu and Yu (2001) has been able to separate repeats differing 5% or more in simulated data, and it remains to be tested on repeats that are more similar.

Another recent method (Pevzner *et al.*, 2001) is based on error correction followed by an Eulerian path approach. Although promising, the error correction step is at most comparable to our basic method, since no statistics are computed to determine the status of positions that occur in a minority. The statistical approach used in our method can probably be used to improve the error correction step of this method, and it would be interesting to see the results of such a combination. The utility of these two other recent approaches remains to be determined, since none of them has yet been rigorously tested.

In comparison, the new method presented in this paper has the statistical power necessary to separate nearly identical repeats. This is due to the use of multi-alignments, the use of statistical calculations in order to extract more information, and evaluation of the number of the true differences for use in the fragment assembly.

The use of this method in assembly programs will facilitate the finishing of repetitive regions and it will in this way extend the use of the shotgun sequencing strategy. This method will be equally useful both for tandem and for dispersed repeat sequences. The assembly of large data sets containing multiple repetitive regions has been shown to be improved by the use of positional information from the sequencing of both ends of clone inserts. We predict that the use of this strategy in combination with the method for repeat separation presented here will provide powerful tools for the assembly of shotgun projects, such as large clones or bacterial genomes. Due to the heterogeneous nature of diploid genomes, it may be difficult to use this method for whole genome shotgun sequencing of such genomes. In such cases, the method needs to be modified by imposing further restrictions on the use of DNPs.

In summary, we have developed a method that makes it possible to separate shotgun reads from nearly identical repeats that is a great improvement over previous methods. The testing of the method shows it is close to the theoretical limit for repeat separation. Work is in progress to incorporate this strategy in an assembly program.

ACKNOWLEDGEMENTS

The authors wish to thank Daniel Nilsson and the reviewers of this paper for valuable input. This study was supported by The Swedish Technology Research Council, Ulf Pettersson and the Beijer Foundation, NIAID (5 U01 AI 45061-02) and the Knut and Alice Wallenberg Foundation.

REFERENCES

Anson,E.L. and Myers,E.W. (1997) ReAligner: a program for refining DNA sequence multi-alignments. *J. Comput. Biol.*, **23**, 262–272.

Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using Phred II. Error probabilities. *Genome Res.*, **8**, 186–194.

Green,P. (1996) http://www.phrap.org/phrap.docs/phrap.html.

Kececioglu,J. and Yu,,J. (2001) Separating repeats in DNA sequence assembly. In *Proceedings of the 5th Annual International Conference on Computational Molecular Biology, April 2001*. ACM Press, Montreal, Canada.

Pevzner,P. A., Tang,H. and Waterman,M.S. (2001) A new approach to fragment assembly in DNA sequencing. In *Proceedings of the 5th Annual International Conference on Computational Molecular Biology, April 2001*. ACM Press, Montreal, Canada.

Ross,S.M. (1988) *A First Course in Probability*. Macmillan, New York.

Tammi,M.T., Arner,E. and Andersson,B. (2001) TRAP: Tandem Repeat Assembly Program, produces improved shotgun assemblies of repetitive sequences. *Computer Methods and Programs in Biomedicine*, in press.