Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

New strategy to detect single nucleotide polymorphisms

Miguel Galves[1,2], José Augusto Amgarten Quitzau[2] and Zanoni Dias[1,2]

[1]Instituto de Computação, Unicamp, Campinas, SP, Brasil
[2]Scylla Bioinformática, Campinas, SP, Brasil
Corresponding author: Z. Dias
E-mail: zanoni@ic.unicamp.br/zanoni@scylla.com.br

Genet. Mol. Res. 5 (1): 143-153 (2006)
Received January 10, 2006
Accepted February 17, 2006
Published March 31, 2006

ABSTRACT. A great effort has been made to identify and map a large set of single nucleotide polymorphisms. The goal is to determine human DNA variants that contribute most significantly to population variation in each trait. Different algorithms and software packages, such as PolyBayes and PolyPhred, have been developed to address this problem. We present strategies to detect single nucleotide polymorphisms, using chromatogram analysis and consensi of multiple aligned sequences. The algorithms were tested using HIV datasets, and the results were compared with those produced by PolyBayes and PolyPhred using the same dataset. Our algorithms produced significantly better results than these two software packages.

Key words: Single nucleotide polymorphism, Chromatogram, Algorithm, Base-calling analysis, Sequence alignment, HIV

INTRODUCTION

A polymorphism can be defined as a set of base pair loci in genomic DNA at which different alleles exist in individuals in some population. To be considered as a polymorphism, the second most frequent allele must have an abundance of 1% or greater (Brookes, 1999).

Single nucleotide polymorphisms (SNPs) are polymorphisms that occur in single-base pair positions in genomic DNA. Single base insertions and deletions are not considered to be SNPs. In coding regions, an SNP can be synonymous or not. In the former case, the new codon produces the same amino acid, thus not affecting the protein product. In the latter case, the non-synonymous SNP (nsSNP) produces a different protein and may affect protein function. One of the major interests in human genome research is to find out whether specific nsSNPs affect human health. In fact, approximately half of the causes of genetic diseases are due to substitutions in amino acids (Cooper et al., 1998).

A concerted effort is currently being made to identify and map a large set of SNPs. The SNP Consortium and the Human Genome Sequencing Consortium published in November 2000 a map containing 1.42 million SNPs of the human genome (Sachidanandam et al., 2001). This interest in SNPs will have a major impact upon population genetics, drug development, cancer, and genetic disease research.

The use of computational tools for SNP analysis is increasing, and sequencing cost is constantly decreasing. Sequencing analysis coupled with appropriate software is a very powerful way to discover new SNPs. One chooses a genomic region of interest and obtains its DNA sequence from different individuals. The sequences that are obtained are aligned, using alignment algorithms (Gotoh, 1982; Myers and Miller, 1998). The final alignment allows sequence comparisons and the detection of possible SNP candidates. The use of additional information such as chromatogram reads can be very useful.

We developed new strategies and algorithms to detect new SNPs, using chromatogram analysis and multiple-sequence consensus construction. The algorithms were tested using HIV data, and the results were compared to those obtained by two publicly available software packages: PolyBayes (Marth et al., 1999) and PolyPhred (Nickerson et al., 1997).

MATERIAL AND METHODS

Our new algorithm to detect new SNPs uses four steps: sequence trimming, sequence chromatogram analysis, polymorphism filtering, and consensus generation.

HIV dataset description

We used HIV genetic sequences, obtained from seropositive individuals. The region corresponds to a well-conserved site composed of 1,302 bp, which appears almost identically in different HIV strains.

The sequences were grouped into 35 batches, each one corresponding to an individual. Each batch contained six PCR reads from the same region, with an average size of 670 bp, along with a validated sequence with a size of 1302 bp with manually mapped polymorphism annotations. These validated sequences were used to benchmark the results. Moreover, a common reference sequence of the 1302-bp region was used to build multiple alignments for each batch.

Using BLAST (Altschul et al., 1990) to compare our reference sequence against complete HIV genomes from a public database (HIV Databases - http://www.hiv.lanl.gov/), we found out that the 50 best hits had an e-value = 0.0, which means that they are statistically identical, with a base similarity always greater than 97% .

Sequence trimming

The trimming algorithm used in this study, similar to the one implemented by phred (Ewing et al., 1998; Ewing and Green , 1998), identifies the maximum score subsequence. In the first step, the quality sequence was converted to an error probability sequence using the phred error formula: $q = -10*\log_{10}(p)$. The algorithm subtracts 0.05 from each error probability and looks for the maximum score subsequence.

Base-calling analysis

This section describes the strategies implemented for polymorphism identification based on the fluorescent dye intensity data from a sequencing machine output combined with the peak positions inferred by the software phred.

We refer to the peak corresponding to the base called by phred as the reference peak and to the other base peaks as polymorphic peaks. Polymorphism identification consists of comparing the reference peak with the polymorphic ones and deciding whether the polymorphic peaks were created by sequencing errors or if the analyzed sequence position corresponds to a polymorphism.

Concerning the chromatogram peaks, we call the position with the highest signal reference point or reference position. The peak extremities are called the start (left extremity) and the end (right extremity) of the peak.

Reference-polymorphic area ratio

The aim of this strategy was to identify polymorphisms based on the reference-polymorphic peak area ratio. This was done by dividing the base-calling task into two phases: peak identification for the four bases and reference-polymorphic ratio comparison for the three possible polymorphic bases.

The first step for the polymorphic peak identification is the definition of the chromatogram area to be scanned. Figure 1 shows how this region is determined: position i corresponds to the trace point where the software phred has identified a reference peak. Positions i - 1 and i + 1 correspond to the preceding and subsequent reference peaks, respectively.

For each reference peak i, the distances $d_1 = d(i; i - 1)$ and $d_2 = d(i; i + 1)$ are calculated. Here, d(a; b) denotes the distance between two peaks. The region to be scanned in search of polymorphic peaks is given by

$$pd1 \le p(x) \le pd2;$$

where

$$pd1 = p(i) - \text{DISTANCE\_PERCENTAGE} \times d2$$

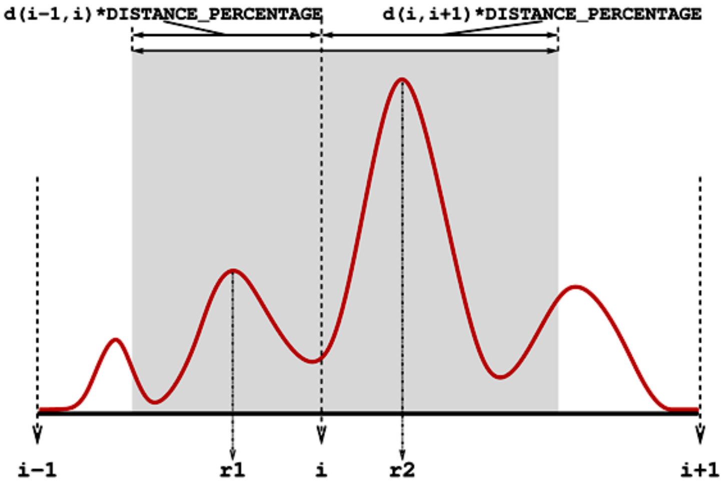

Figure 1. Chromatogram region where the search for polymorphic peaks is performed. The curve represents a secondary trace in the chromatogram, not taken by phred as the called base in the given position. The positions i - 1, i and i + 1 correspond to the positions of consecutive reference peaks. The function d(a; b) determines the distance between two peaks. In this example, only the peaks whose reference positions are in the shaded area (r1 and r2) would be considered in the search for polymorphisms.

and

$$pd2 = p(i) + DISTANCE\_PERCENTAGE \times d1,$$

and p(x) denotes the chromatogram point where the xth reference position is found and DISTANCE PERCENTAGE is a parameter such that

$$0 \leq DISTANCE\_PERCENTAGE \leq 1$$

The algorithm scans the region and identifies chromatogram points that satisfy the condition:

$$t(p - 1) \leq t(p) > t(p + 1)$$

or

$$t(p - 1) < t(p) \geq t(p + 1);$$

where t(p) corresponds to the intensity of the fluorescent dye observed at the pth scan, that is, at the pth chromatogram point. All points found are considered “peaks”. However, only the peaks nearest to the positions given by phred are considered as possible polymorphisms.

The analysis starts with the determination of the region that corresponds to the peak. Initially, both the start and end point of the peak region are equal to the reference point. In the first step, these points are moved apart from the reference point, while the values at the chromatogram points to the left of the start and at the right of the end points are less than the values at the start and end points, respectively. At the end of this step, the region extremities correspond to a local minimum or to the beginning of a plateau. In the case of plateaus, the entire flat region is identified, and the plateau midpoint is considered as the region limit.

Using Figure 2 as an example, the first identified point is the peak reference, which corresponds to its maximum. After this, the start and end points are defined. In the case illustrated by this example, the end point determined at this step is found at the point labeled plateau start point. In the third step, this point is moved to the point labeled plateau end point and the plateau midpoint is used as the peak end point.

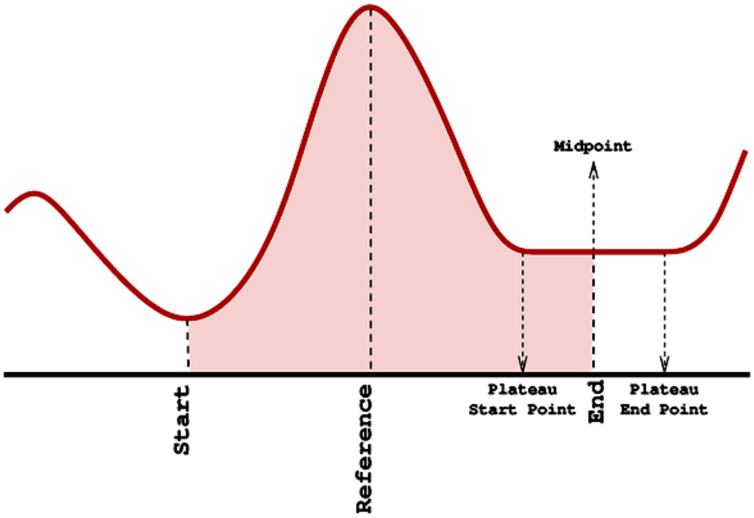

Figure 2. Limits considered during peak analysis. The shaded region corresponds to the scanned region. The point labeled Reference indicates the chromatogram point used as a reference for the definition of the scanned region. The points Start and End correspond to the peak limits in the chromatogram. Note that the position End corresponds to the midpoint of a plateau.

The peak identification procedure is repeated for each of the four nucleotides. After this, the peak area of the base identified by phred is compared to the largest peak area between the polymorphic bases. This is done by the analysis of the reference-polymorphic peak area ratio. If this ratio is greater than the limit established by the parameter MIN_RELATION, then the position analyzed is considered a polymorphism and the original base is changed to one of the standard IUPAC symbols (International Union of Pure and Applied Chemistry - http://www.chem.qmw.ac.uk/iupac/) for base pairs.

Reference-polymorphic average height ratio

Like the area-based ratio, this strategy was also divided into two steps: identification of peaks and comparison between the fluorescent dye intensity of the base identified by Phred and the intensity of the other bases.

The peak identification in this strategy is almost the same as the peak identification in the previous strategy. The only difference is that this strategy ignores very small peaks that correspond to natural small variations in dye intensity, which are not peaks. This fake peak identification is based on two parameters: FAKE_PEAK_HEIGHT_PERCENTAGE and MAXIMUM_FAKE_PEAK_WIDTH, which determine the limits for height and width of a fake peak.

Figure 3 illustrates the effect of variation on these parameters. In this strategy, when a peak extremity cannot be extended, the algorithm tries to identify a fake peak. A fake peak is defined as a region with no more than MAXIMUM_FAKE_PEAK_WIDTH trace points, where the highest trace value is not greater than FAKE_PEAK_HEIGHT_PERCENTAGE percent of the value at the nearest extremity of the considered peak.

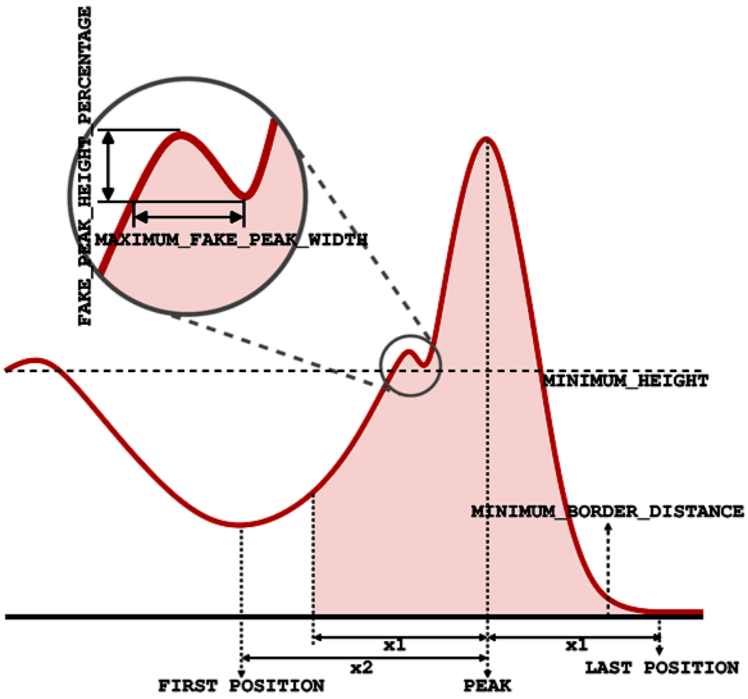

Figure 3. Peak identified by the average height ratio method. A fake peak was amplified so that the parameters used to identify it are made more visible.

Like the previous strategy, the scanned region is defined by the parameter DISTANCE_PERCENTAGE. Nevertheless, in this case the nearest peak may not be analyzed when it does not satisfy the following conditions:

Minimum extremity distance: Both peak extremities must be at a certain distance from the reference position. This distance must be greater than a minimum distance determined by the parameter MINIMUM_BORDER_DISTANCE.

Minimum height: The maximum fluorescent dye intensity must be at least the value established by the parameter MINIMUM_HEIGHT.

There are basically two differences in the way the two strategies analyze the peaks. First, it is not the area but the average height. Thus, the compared value is the average intensity of the trace points in the peak. The second difference is that the peak reference position is centralized. This is done by changing the start or end position in such a way that both the distance between the start and the reference and the distance between the reference and the end points are the same. This distance is determined by the minimum distance between these two pairs of points. Figure 3 shows this position change. We can see that the area used for average height calculation does not start at the peak extremity, but at the position with distance equal to x1 from the reference.

The average heights are compared in the same way as the areas are compared in the first strategy. Initial default parameter values for reference-polymorphic area ratio strategy are MIN_RELATION = 0.3 and DISTANCE_PERCENTAGE = 0.2. For reference-polymorphic average height ratio strategy, the initial default parameter values are MIN_RELATION = 0.15, DISTANCE_PERCENTAGE = 0.3, FAKE_PEAK_HEIGHT_PERCENTAGE = 0.1, MAXIMUM_FAKE_PEAK_WIDTH = 5, MINIMUM_BORDER_DISTANCE = 7, and MINIMUM_HEIGHT = 50.

Polymorphism filter

To reduce the number of false-positives generated by the chromatogram analysis step, described above, we implemented a filter that analyzes SNP distribution along a sequence. The main goal here was to avoid sequence segments with adjacent polymorphisms, probably generated by sequencing errors.

Consensus algorithm

The consensus algorithm that is implemented is a rule-based algorithm that aims at generating a consensus analyzing all the bases in a cross-section of the multiple alignment result. The rules were defined empirically.

First of all, the algorithm counts how many times each different symbol appears in the cross-section, and it calculates the average quality. If polymorphism symbols appear in the section, they are counted as separated symbols and the bases represented by each symbol are also incremented. For instance, if the cross-section is composed of the symbols A, A and M, with qualities 30, 20 and 10, then the algorithm reports three A's with an average quality of 20, one M with an average quality 10 and one C with an average quality of 10 (M = A + C). Once the frequency of each symbol and its corresponding average quality has been determined, the following rules are applied:

- If only one symbol is reported in a cross-section, then the consensus is this symbol;
- If no polymorphism is accounted for, but more than one type of base exists in the cross-section, the base with the highest frequency should be examined to check whether it has an average less than or equal to 20, whether its average quality is 50% higher than the second base, and whether it appears at least three times more than the second base. In case such conditions hold true, the algorithm returns the IUPAC symbol corresponding to a polymorphism. Otherwise, the algorithm returns the most frequent base.

- If there is a defined polymorphism in the base-calling phase represented by a corresponding IUPAC symbol in the cross-section, we check whether it is compatible with the two most frequent bases. If it is not, the most frequent base is the consensus. Otherwise, we check whether the polymorphism occurs in at least 2/3 of the cross-section bases or whether the average quality of the polymorphism is higher than 50% of the quality of the most frequent base; in case of any one of these conditions, the consensus is polymorphism, otherwise the consensus is the most frequent base.

Table 1 shows the consensus sequence generated by the algorithm described above, given a set of aligned sequences.

Table 1. Consensus sequence generated by the consensus algorithm. The first four lines represent the considered sequences, with qualities in subscript.

Sequence 1	A_{25}	C_{30}	C_{18}	C_{30}	A_{21}
Sequence 2	A_{30}	C_{25}	C_{15}	C_{25}	A_{16}
Sequence 3	-	M_{18}	A_{09}	C_{30}	-
Sequence 4	-	-	S_{12}	G_{17}	T_{18}
Consensus	A	M	S	S	W

RESULTS

Figures 4, 5 and 6 show plots of comparisons between results obtained using our strategies with certain parameters, and those obtained by PolyPhred and PolyBayes, with the same parameters.

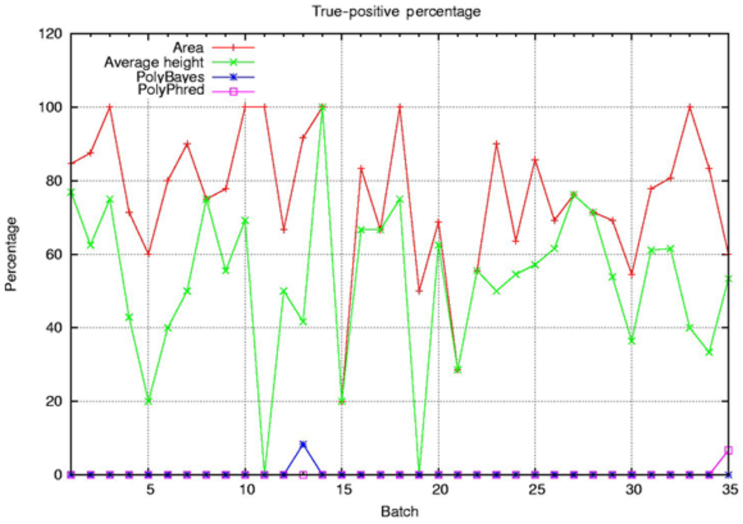

Figure 4. True-positive average percentage (reported/expected × 100) using PolyBayes, PolyPhred, reference-polymorphic area ratio and reference-polymorphic average height ratio strategies, by batch.

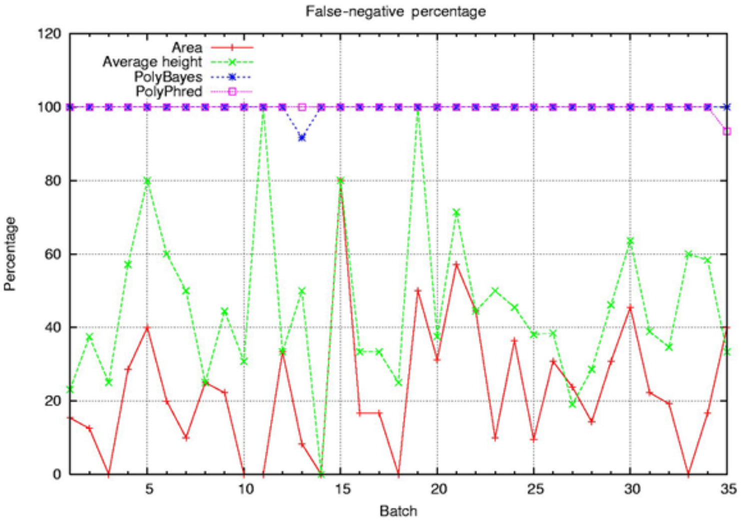

Figure 5. False-negative average percentage (reported/expected × 100) using PolyBayes, PolyPhred, reference-polymorphic area ratio and reference-polymorphic average height ratio strategies, by batch.

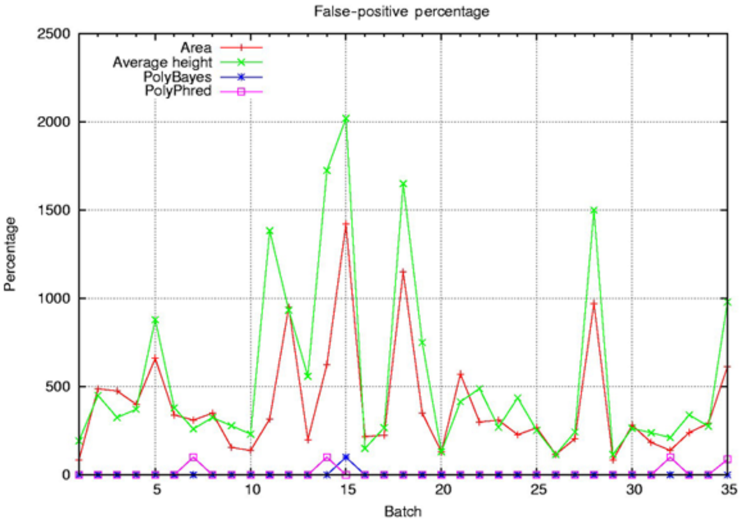

Figure 6. False-positive average percentage (reported/expected × 100) using PolyBayes, PolyPhred, reference-polymorphic area ratio and reference-polymorphic average height ratio strategies, by batch.

To run tests, the following steps were taken: we used the reference sequence as anchor to generate a contig and a consensus sequence with the application phrap (Gotoh, 1982). The generated ACE file was processed by the software packages PolyBayes and PolyPhred, which created an output file with all identified polymorphisms. To compare the resulting consensus with the validated sequence of each batch, cross_match (Gotoh, 1982) was used to generate the alignment.

Of 35 batches with all marked polymorphisms, PolyBayes detected polymorphisms in only two batches, and PolyBayes detected polymorphisms in four batches.

The average size of sequences, before the process of removing low-quality regions, was 691.15 bp. After filtering, sequences were shortened to an average size of 374.74 bp, showing an average reduction in sequence size of 45%. The average base coverage using original sequences (without the trimming process) was 2.69 reads, with a maximum value of 2.92 reads, a minimum value of 2.15 reads, and a standard deviation of 0.14 reads. All bases in the reference sequence are covered by the final consensus sequence based on the original sequences. Using trimmed sequences, the average base coverage was 1.77 reads, with a maximum value of 2.51 reads, a minimum value of 1.24 reads, and a standard deviation of 0.31 reads. On average, 5.16% of the reference sequence's bases were not covered by consensus, based on trimmed sequences. Average coverage was obtained by summing up the total number of bases covering a given position in the reference sequence, and dividing the results by the number of covered bases in the reference (Table 2).

Table 2. Summary (average percentage (Avg.) and standard deviation(SD)) of true-positives (TP), false-negatives (FN), false-positives (FP), and different polymorphisms (DP - detected polymorphism is different from a polymorphism that is manually annotated) obtained by PolyPhred, PolyBayes, reference-polymorphic area ratio and reference-polymorphic average height ratio base-calling algorithms (TP + FN + DP = 100%).

	PolyBayes		PolyPhred		Area		Avg. height	
	Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD
True-positives (%)	0.2	1.4	0.2	1.1	75.4	19.2	52.6	21.5
False-negatives (%)	99.8	1.4	99.8	1.1	23.2	18.4	45.6	21.7
Different polymorphisms (%)	0.0	0.0	0.0	0.0	1.4	4.3	1.8	4.0
False-positives (%)	2.9	16.9	11.1	31.3	393.9	312.3	554.4	511.3

DISCUSSION

The results obtained by PolyBayes and PolyPhred packages were far from acceptable, when applied to the data set that we used here. In comparison, the results obtained by our algorithms were quite satisfactory, taking into account the type of batch used: the low average coverage of each base in the reference sequence, the high percentage of low-quality bases and the great amount of polymorphism.

Clearly, the reference-polymorphic area ratio strategy has a better performance than the reference-polymorphic average height ratio. We varied the parameters, and determined that the best values for our algorithm are MIN_RELATION = 0.25 and DISTANCE_PERCENTAGE = 0.5. The former strategy was as efficient as or more efficient than the other strategy in 100% of the cases when compared percentage-wise with regard to the number of correct polymorphisms (Figure 4). The graph shown in Figure 5 clearly shows the comparative efficiency of the first strategy with regard to false-negatives. Only when compared percentage-wise with respect to false-positives found, was the efficiency of using areas found to be not so evident (Figure 6). Even so, we can see that in only 12 batches, approximately 34% of the total, this strategy was outperformed by the height average strategy.

It would be useful to run this experiment with higher quality sequences and larger batches, so that it can provide a better coverage and allow a better choice of parameters for the SNP filter and consensus algorithms. Furthermore, viruses such as the HIV used in this study have a large number of mutations. It would be interesting to reproduce the study using genetic sequences of more conserved life forms such as mammals, for instance, to validate the algorithms that were developed.

ACKNOWLEDGMENTS

The present study was developed at Scylla Bioinformatics (www.scylla.com.br) and was partially financed by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo), grants 2003/07748-9 and 2004/09418-9. We thank João Meidanis for manuscript revision.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, et al. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.

Brookes AJ (1999). The essence of SNPs. *Gene* 234: 177-186.

Cooper DN, Ball EV and Krawczak M (1998). The human gene mutation database. *Nucleic Acids Res.* 26: 285-287.

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

Ewing B, Hillier L, Wendl MC and Green P (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8: 175-185.

Gotoh O (1982). An improved algorithm for matching biological sequences. *J. Mol. Biol.* 162: 705-708.

Green P (2002). Phrap documentation, September 2002. www.phrap.org.

Marth GT, Korf I, Yandell MD, Yeh RT, et al. (1999). A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* 23: 452-456.

Myers EW and Miller W (1988). Optimal alignments in linear space. *Comput. Appl. Biosci.* 4: 11-17.

Nickerson DA, Tobe VO and Taylor SL (1997). PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* 25: 2745-2751.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409: 928-933.