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

Analysis of slipped sequences in EST projects

Christian Baudet[1,2] 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): 169-181 (2006)
Received January 10, 2006
Accepted February 17, 2006
Published March 31, 2006

ABSTRACT. Slippage is an important sequencing problem that can occur in EST projects. However, very few studies have addressed this. We propose three new methods to detect slippage artifacts: arithmetic mean method, geometric mean method, and echo coverage method. Each method is simple and has two different strategies for processing sequences: suffix and subsequence. Using the 291,689 EST sequences produced in the SUCEST project, we performed comparative tests between our proposed methods and the SUCEST method. The subsequence strategy is better than the suffix strategy, because it is not anchored at the end of the sequence, so it is more flexible to find slippage at the beginning of the EST. In a comparison with the SUCEST method, the advantage of our methods is that they do not discard the majority of the sequences marked as slippage, but instead only remove the slipped artifact from the sequence. Based on our tests the echo coverage method with subsequence strategy shows the best compromise between slippage detection and ease of calibration.

Key words: Slippage, Expressed sequence tag, Sequence trimming, Sugar cane

INTRODUCTION

The objective of expressed sequence tag (EST) sequencing projects is to quickly obtain the gene index of an organism, which is the set of all genes that exist in the genome of an organism. An EST (Adams et al., 1991) is a cDNA (complementary DNA), that is a copy of an mRNA (messenger RNA). By sequencing a cDNA, we obtain a nucleotide sequence belonging to a gene that exists in the genome and that is expressed by the cell.

The EST sequencing process includes cDNA library production, cDNA cloning, and clone sequencing. The last step is processed by a single run in a sequencing machine, which is one of the characteristics that makes this technique more cost-effective than the other ones. The chromatograms produced by these sequencing machines are processed by base-calling software that determines the base sequence of the EST. This software usually produces quality values for each base. The quality value indicates the probability of error of the call. Usually EST sequences have artifacts, such as low quality regions, poly-A/T tails, vector, and adapter sequences. These artifacts must be removed because they can spoil the data analysis. Thus, sequence trimming is an important step that must be performed in EST sequencing projects.

Slippage is an artifact type that can be found in EST sequences. Caused by problems in the sequencing process, slippage is a region that has an abnormal distribution of echoed bases. These echoes result from reading chromatogram regions that have many signal peaks for a single nucleotide.

In cDNA sequences, slippage is related to long poly-A/T tails. Long tails may have problems in staying paired during the polymerization reaction and this can generate fragments with homopolymer regions of different lengths that have the same sequence after the region (Applied Biosystems, 1998).

Although echoed bases sometimes appear with a high background noise, signal peaks are so high that base-calling softwares assign high quality values to bases that do not exist. This phenomenon prevents the removal of these regions by trimming methods based on quality. We developed a set of trimming procedures for EST sequencing projects (Baudet and Dias, 2005).

During our research, we observed that only Telles and da Silva (2001) had carried out a study of slipped sequences. Their method defines an echoed region as a set of at least five identical consecutive bases. The product of echoed region lengths is evaluated for each sequence. If the echoed region length is equal to or greater than 10, it contributes just 10 to the product. Sequences with product greater than 108 and echoed regions covering more than 20% of sequence length are considered slipped.

Once a sequence is considered slipped, an additional step, which searches for poly-A/T tails, is performed to define the subsequence that will be marked as a slippage artifact. If a poly-T is found, the whole sequence is discarded because the tail is usually placed at the 5’ end. If a poly-A, which is usually placed at the 3’ end, is found, only its own sequence and the remaining 3’ sequence are discarded. If nothing is found, the whole sequence is discarded.

The method above imposes a minimum coverage of 20%. This can be a problem as sequence length grows. If a sequence has 600 bases and the slipped region has a length of less than 120 bases, correct artifact identification will not happen if we use these criteria. Moreover, we observed that this method does not demand proximity of the echoed regions. With the goal of improving slippage detection, we examined three new alternatives and compared them with the existing method.

MATERIAL AND METHODS

The three slippage-detection methods proposed in this study are simple. Each one of the methods has two strategies on how to process a sequence.

The first strategy processes the sequence from its end backwards to find the largest suffix that reaches the threshold value. This strategy, which will be called suffix, assumes that slippage affects all bases from its initial position up to the sequence end.

The second strategy, called subsequence, performs the search of maximal subsequences that have scores greater than the threshold value. This strategy considers that slippage has its start and end positions clearly defined, and that it is possible to discover them. Thus, the remaining sequence, which is not slipped, can be used in other analyses.

These methods also have two common parameters. In the sequel, a group is a set of one or more identical consecutive bases. We define the parameters as follows:

- *minimum_echo_size* defines the minimum length that a group must have to be considered as an echoed group.
- *minimum_number_of_echoes* defines the minimum number of echoed groups that a region must have to be considered for analysis.

An important detail that should be noted is that all methods consider as valid echoed groups those that are only composed of bases A, T, C, or G. Groups formed by Ns are not considered as echoed groups because they are, in fact, low quality artifacts. They can produce negative effects in the scores calculated by the methods and may point to slippage artifacts that do not exist.

Method 1 - Arithmetic mean

This method calculates, for a region, the ratio between the sum of all echoed group lengths and the total number of groups.

If we use this method with the parameters *minimum_number_of_echoes* = 3 and *minimum_echo_size* = 4 and suffix strategy, the sequence

A	T	C	G	TTTTTT	AAAAA	CCC	GGGGG	TT	CCC	AAAA	TT
1	1	1	1	6	5	3	5	2	3	4	2

produces the following suffixes

AAAAACCCGGGGGTTCCCAAAATT	(4 + 5 + 5)/7 = 2.00
TTTTTTAAAAACCCGGGGGTTCCCAAAATT	(4 + 5 + 5 + 6)/8 = 2.50
GTTTTTTAAAAACCCGGGGGTTCCCAAAATT	(4 + 5 + 5 + 6)/9 = 2.22
CGTTTTTTAAAAACCCGGGGGTTCCCAAAATT	(4 + 5 + 5 + 6)/10 = 2.00
TCGTTTTTTAAAAACCCGGGGGTTCCCAAAATT	(4 + 5 + 5 + 6)/11 = 1.81
ATCGTTTTTTAAAAACCCGGGGGTTCCCAAAATT	(4 + 5 + 5 + 6)/12 = 1.67

In this case, the best suffix has the score 2.50. If the same sequence is analyzed with the same parameters and subsequence strategy, the region

TTTTTTAAAAACCCGGGGGTTCCCAAAATT $(4 + 5 + 5 + 6)/7 = 2.86$

is identified as the best subsequence.

Method 2 - Geometric mean

The geometric mean method is similar to the previous method. The difference is that it calculates the region score as the product of echoed group lengths raised to the inverse of the number of groups. Thus, the suffix scores calculated for the same sequence above are

AAAAACCCGGGGGTTCCCAAAATT	$(4 \times 5 \times 5)^{1/7} = 1.93$
TTTTTTAAAAACCCGGGGGTTCCCAAAATT	$(4 \times 5 \times 5 \times 6)^{1/8} = 2.22$
GTTTTTTAAAAACCCGGGGGTTCCCAAAATT	$(4 \times 5 \times 5 \times 6)^{1/9} = 2.04$
CGTTTTTTAAAAACCCGGGGGTTCCCAAAATT	$(4 \times 5 \times 5 \times 6)^{1/10} = 1.90$
TCGTTTTTTAAAAACCCGGGGGTTCCCAAAATT	$(4 \times 5 \times 5 \times 6)^{1/11} = 1.79$
ATCGTTTTTTAAAAACCCGGGGGTTCCCAAAATT	$(4 \times 5 \times 5 \times 6)^{1/12} = 1.70$

Method 3 - Echo coverage

This method applies a transformation to the sequence to evaluate the echoed group coverage of the analyzed region. In this transformation, every echoed group is replaced by 1 and every normal group is replaced by 0. The transformation of our sequence example is 000011010010.

After transformation, the method calculates the ratio between the number of 1s and the transformed region length. Consequently, the suffixes of our example have the following scores:

AAAAACCCGGGGGTTCCCAAAATT	1010010	3/7 = 0.43
TTTTTTAAAAACCCGGGGGTTCCCAAAATT	11010010	4/8 = 0.50
GTTTTTTAAAAACCCGGGGGTTCCCAAAATT	011010010	4/9 = 0.44
CGTTTTTTAAAAACCCGGGGGTTCCCAAAATT	0011010010	4/10 = 0.40
TCGTTTTTTAAAAACCCGGGGGTTCCCAAAATT	00011010010	4/11 = 0.36
ATCGTTTTTTAAAAACCCGGGGGTTCCCAAAATT	00011010010	4/12 = 0.33

To perform tests with these methods, we employed the same data set used by Telles and da Silva (2001). This data set is composed of 291,689 sugar cane EST sequences from the SUCEST project (Vettore et al., 2001). The average sequence length is 829.44 ± 182.60 bases. The average number of bases with PHRED (version 0.980904.e) (Green, 2005) quality greater than 20 is 399.53 ± 182.60 bases. The sequences were produced from 26 libraries. Most of the sequences (259,325) were sequenced from the 5' end, while the remaining sequences (32,364) were sequenced from the 3' end. We also implemented the SUCEST method for comparison purposes.

In our tests, we performed BLAST (version 2.2.11) (Altschul et al., 1997) of slipped sequences against the Swiss-Prot database (release 46.6 - April 26, 2005) (Boeckmann et al.,

2003) to determine the influence of slippage removal in the gene detection process. Each slipped sequence was compared against the Swiss-Prot database in three different ways: i. Complete sequence with no masking; ii. complete sequence with vector masking, to measure the approximate number of hits found in the previous manner due to vector fragments; iii. largest contiguous subsequence that was not masked as slippage or as vector. Sequences with a length of less than 100 bases were discarded.

The vector masking of the sequences was performed through the execution of cross_match (version 0.990319) (Green, 2005) using the parameters -*minmatch 12* and -*minscore 20*. The slippage masking for our proposed methods masked the longest slipped sequence that had a score equal to or greater than the threshold score that was chosen for the method. The slippage masking for the SUCEST method was performed as described in their work.

For each pair method/strategy and for the SUCEST method, we observed the percentage of sequences with at least one hit with e-value equal to or less than 10-5. This e-value was selected because the Swiss-Prot database is very well curated. All methods were implemented in Perl (version 5.8.5) (CPAN, 2005).

RESULTS

Identifying the suffix or subsequence with the highest score does not necessarily mean finding the whole slippage. Depending on echoed group distribution, the slippage score can be lower than the score of a subsequence that is contained in it. Therefore, the proposed methods need the definition of threshold values. The slipped region would be the largest suffix or subsequence that has a score equal to or greater than the method threshold value.

The first step in our tests was the execution of each one of the two strategies of each method with *minimum_echo_size* varying in the interval {1, 2, ... , 10}. The parameter *minimum_number_of_echoes* was set to 8 for comparison purposes because this is the minimum number of echoed groups that is necessary to reach 10^8 in the SUCEST method. For each execution, one list was produced with the maximum scores for the suffix or subsequence of each sequence in the data set.

Since data volume was very high, each list was sorted in ascending order and divided into 100-sequence intervals. The mean score of each interval was then calculated. Figures 1, 2 and 3 show the surface graphs made with the results of arithmetic mean, geometric mean, and echo coverage methods, respectively, running with the suffix strategy. The subsequence strategy produced graphs with similar behavior.

To better illustrate the behavior from distinct method/strategy pairs, we selected, for each one, a different base value as its slippage score threshold. We counted the number of sequences with scores greater than or equal to the threshold value. We repeated this procedure varying the threshold value by adding -15, -10, -5, -2, -1, 1, 2, 5, 10, and 15% to the original base value. The result of this procedure for the pairs method/strategy with *minimum_echo_size* = 5 is shown by the graph in Figure 4.

The second step of our tests was to compare the results of our method results with the sequences reported as slipped by the SUCEST method. This test was performed to evaluate the detection capacity through the contrast of different method results.

We implemented the SUCEST method (Method 4) according to the description found in their work. This implementation was used to process the same data set and 7213 sequences

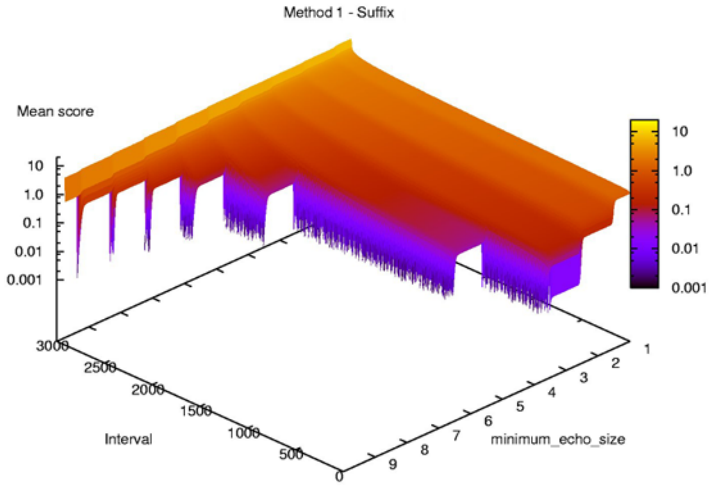

Figure 1. Arithmetic mean method executed with *minimum_number_of_echoes* = 8, *minimum_echo_size* = [1, 10], and suffix strategy. The results of each execution were sorted in ascending order and split in 100-sequence intervals, then their mean score was calculated. This graph shows the behavior of these intervals in each execution. The mean score axis is in logarithmic scale.

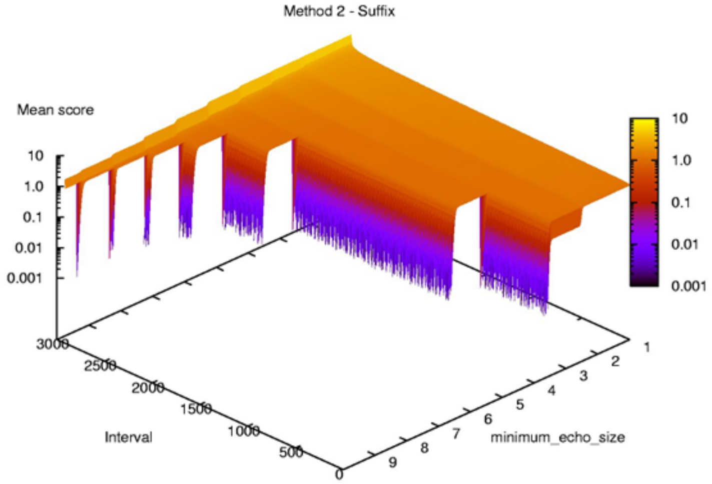

Figure 2. Geometric mean method executed with *minimum_number_of_echoes* = 8, *minimum_echo_size* = [1, 10], and suffix strategy. The results of each execution were sorted in ascending order and split in 100-sequence intervals, then their mean score was calculated. This graph shows the behavior of these intervals in each execution. The mean score axis is in logarithmic scale.

were marked as slippage. The processing was carried out with raw sequences and not with partially trimmed sequences, as in their work.

This comparison was performed with the results of all methods using the value 5 for the parameter *minimum_echo_size*. This was the same value adopted by Telles and da Silva (2001)

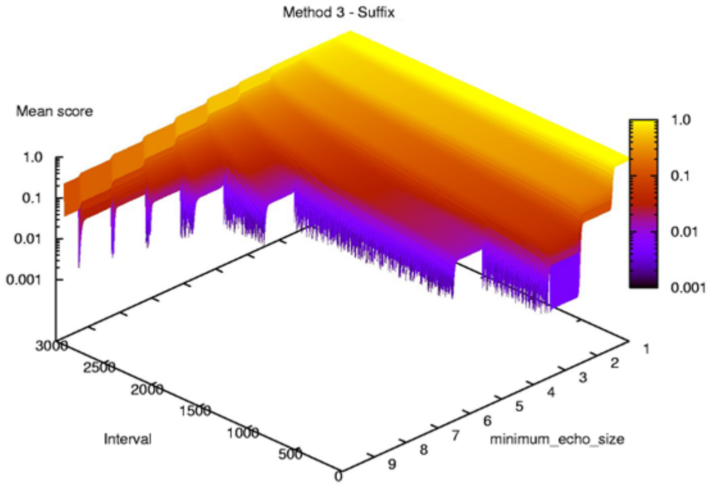

Figure 3. Echo coverage method executed with *minimum_number_of_echoes* = 8, *minimum_echo_size* = [1, 10], and suffix strategy. The results of each execution were sorted in ascending order and split in 100-sequence intervals, then their mean score was calculated. This graph shows the behavior of these intervals in each execution. The mean score axis is in logarithmic scale.

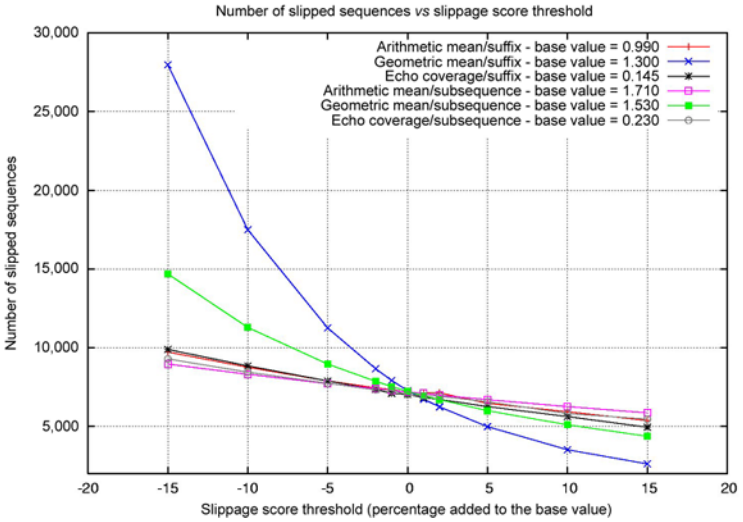

Figure 4. Number of sequences marked as slipped for each method/strategy pair using *minimum_echo_size* = 5. A base value was selected for each pair as the slippage score threshold. The threshold was varied by adding -15, -10, -5, -2, -1, 1, 2, 5, 10, and 15% to the original base value.

and it looks appropriate because it does not restrict the detection of slippage that does not have large echoed groups.

For each pair method/strategy, we selected the 7213 sequence with the highest score. Venn-Euler diagrams were constructed (Figures 5 and 6). They show the intersection of the

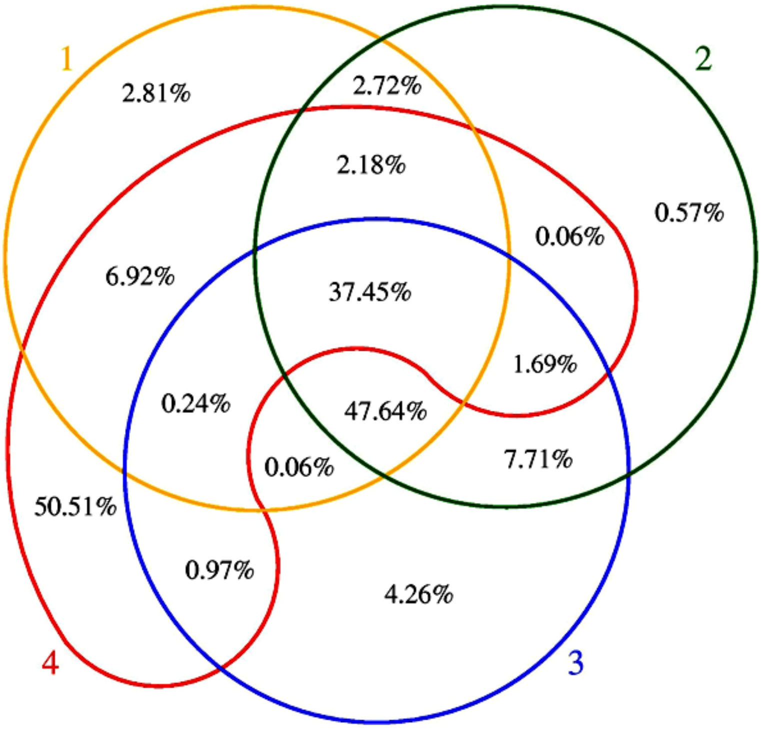

Figure 5. Venn-Euler diagram showing intersections of arithmetic mean method (1), geometric mean method (2), and echo coverage method (3) sets using suffix strategy with *minimum_echo_size* = 5 and the SUCEST method (4) set. Each percentage value indicates the percentage of sequences of the method set that are inside the associated region. For example, the percentage value 0.97% indicates that only this percentage of sequences of the method 3 set are in the set composed of sequences that are only in the method 3 set and in the method 4 set. As all sets have the same size (7213 sequences), the same observation can be made for method 4 for the same region in the diagram.

sequence sets built by each proposed method and by the SUCEST method, respectively, for suffix and subsequence strategies.

The lists of 7213 sequences with the highest scores were also used in the third step of our tests. In this step we compared the strategy pairs of each method. The intersection set of each pair had a size of 4976, 4969 and 4922 for the arithmetic mean, geometric mean and echo coverage methods, respectively.

We sorted each list of each pair in descending order by score. We divided them into intervals of 200 sequences, and for each interval we counted the number of sequences in it that were not found in the sequence set of the other strategy. The two strategies were then compared (Figure 7).

The last test that we performed was the BLAST, as described in the previous section, with the 7213 sequences marked as slipped by each of the proposed method/strategy pairs and by the SUCEST method. Our objective was to determine the influence of slippage removal in gene detection. Table 1 shows the results of the BLAST runs and the percentage loss of hits when we compared the results of the first and second procedures and the results of the second and third procedures.

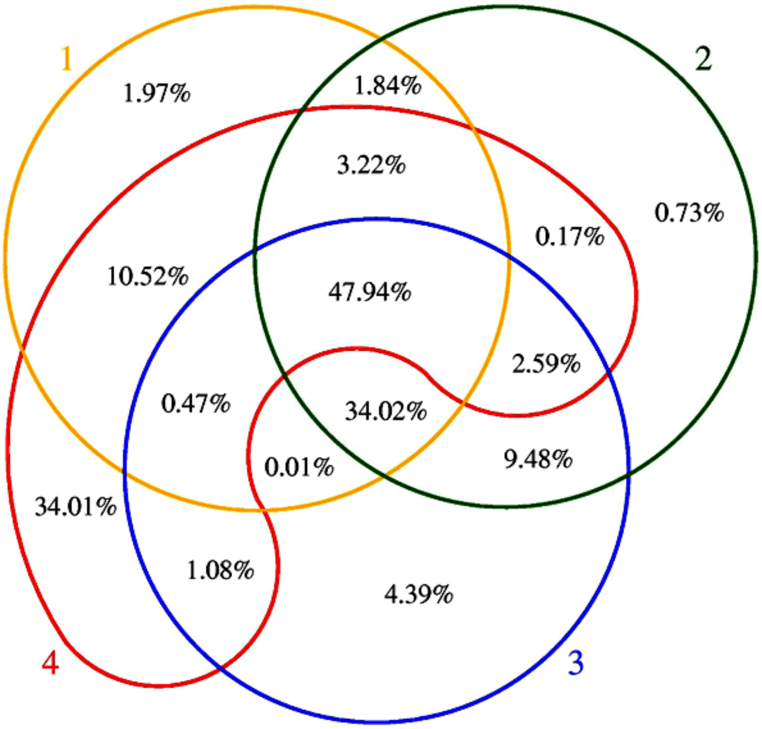

Figure 6. Venn-Euler diagram showing intersections of arithmetic mean method (1), geometric mean method (2), and echo coverage method (3) sets using subsequence strategy with *minimum_echo_size* = 5 and the SUCEST method (4) set. Each percentage value indicates the percentage of sequences of the method set that are inside the associated region. For example, the percentage value 9.48% indicates that only this percentage of sequences of the method 3 set are in the set composed of sequences that are only in the method 2 set and in the method 3 set. As all sets have the same size (7213 sequences), the same observation can be made for method 2 for the same region in the diagram.

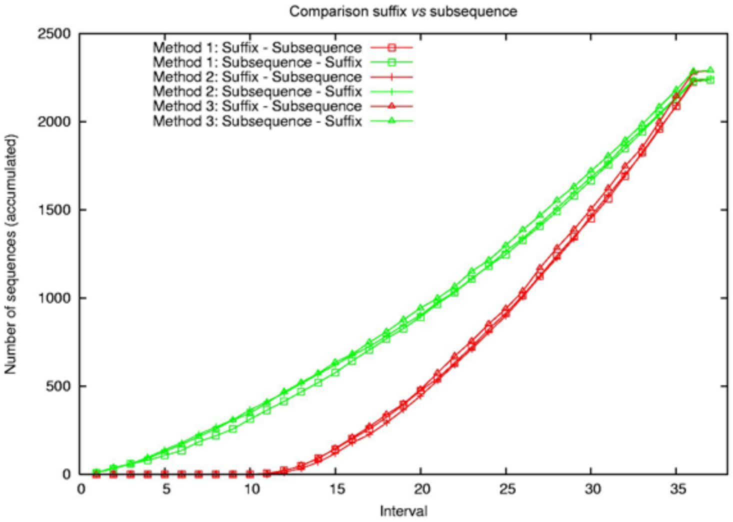

Figure 7. Comparison between the strategy pairs of each proposed method. Each result list was sorted in descending order. The first intervals have the sequences with greatest scores for each method/strategy. The green lines show the accumulated number of sequences that are in a subsequence set's interval and are not in the suffix strategy set. The red lines show the accumulated number of sequences that are in a suffix set's interval and are not in the subsequence strategy set.

Table 1. Number of sequences with at least one hit with e-value lower than or equal to 10^{-5} in each one of the sets of 7213 sequences marked as slipped by the proposed method (I - Arithmetic mean, II - Geometric mean, or III - Echo coverage/suffix or subsequence, and IV - the SUCEST method) with minimum echo size = 5. Each set of sequences was compared against the Swiss-Prot database in 3 different manners: i - complete sequence, ii - complete sequence with vector masking, and iii - largest contiguous subsequence with vector and slippage masking. The last two columns show the percentage of lost hits when we compare the results of the first and second manners and the results of the second and third manners, respectively.

Method	Strategy	i	ii	iii	(1 - (ii/i))	(1 - (iii/ii))
I	Suffix	2532	2034	1946	19.67%	4.33%
II		2811	2225	2166	20.85%	2.65%
III		2856	2211	2147	22.58%	2.89%
I	Subsequence	1938	1590	1516	17.96%	4.65%
II		2198	1781	1716	18.97%	3.65%
III		2287	1831	1765	19.94%	3.60%
IV	-	1443	710	85	50.08%	88.03%

DISCUSSION

Figures 1, 2, and 3 show the behavior of the score distribution of the suffix strategy for the arithmetic mean, geometric mean, and echo coverage methods, respectively. The score distribution of the subsequence strategy was similar, despite the score increase caused by the shorter length of the slipped regions that it detected.

These graphs indicate that all the methods show a similar behavior as we change the value of the *minimum_echo_size* parameter. The bands formed by intervals that have a zero mean score are the same in all graphs. A zero mean score means that every sequence in the interval does not have the minimum number of echoed groups (eight in our tests) with a length greater than or equal to the *minimum_echo_size*. If a sequence does not have a suffix that meets the criterion, it also does not have a subsequence, and vice versa. For example, if we examine the results of the echo coverage method executed with *minimum_echo_size* = 5, we see that only 618 intervals have a non-zero mean score.

Higher scores indicate a higher probability of slippage. When we observe the behavior over the intervals, we can see that the number of high-scoring intervals is very small, compared to the total number of intervals. The variation between intervals is more conserved in the geometric mean method. The graph shows little variation and evidence that this method will be difficult to calibrate, because small variations in the threshold value include or exclude a high number of sequences.

In examining Figure 4, we can confirm this hypothesis. The geometric mean method's curves grow quickly when we diminish the threshold value, which does not happen for the other two methods. An inspection of the arithmetic mean and echo coverage methods shows that their curves are practically linear with both strategies. Therefore, the arithmetic mean method is easier to calibrate, because its curves have a lower inclination, i.e., the difference in the number of sequences between two threshold values is smaller than in the other methods.

Based on the surface graphs, we decided to use *minimum_echo_size* = 5 for all other tests. As we can see in these graphs, this is the first value that has the capacity of discarding a great number of sequences (approximately 80% of the sequences were directly discarded). Moreover, we considered that this value would allow the detection of slippage that does not have large echoed groups.

For each one of the six method/strategy pairs, we took the 7213 sequences with the highest score. This operation is equivalent to defining, for the arithmetic mean, geometric mean, and echo coverage methods, the threshold values 0.9860, 1.3010 and 0.1429 (suffix strategy) and 1.7070, 1.5306 and 0.2286 (subsequence strategy), respectively.

When we analyze the Venn-Euler diagrams shown in Figures 5 and 6, we see that the intersection of the three proposed methods for suffix strategy (85.09%) is greater than the intersection for subsequence strategy (81.96%). This occurs because the suffix strategy could not detect slippage that occurs in the beginning of the sequence. In these cases, this strategy produces low scores because of the presence of a region that is not slipped in the end of the sequence. Therefore, when the three methods are run with this strategy, they stay anchored in the same suffixes.

We can also see that the geometric mean method is virtually covered by the arithmetic mean and echo coverage methods. Less than 1% of the sequences of its set were identified only by it. This subset is much smaller than the subsets of sequences identified only by the arithmetic mean method (~11%) or only by echo coverage method (~5%).

When we include the results of the SUCEST method in the analyses, we see that the size of the intersection of all sets is smaller for the suffix strategy (37.45%) than for the subsequence strategy (47.49%). The SUCEST method yields results closer to the subsequence strategy probably because, by not being anchored to the sequence end, this strategy has more flexibility into finding the echoed regions. Since the SUCEST method has the same characteristics, it is expected to show results more similar to the subsequence strategy than to the suffix strategy.

The intersections between the suffix set and the subsequence set, for each one of the methods, have 68.99, 68.89 and 68.24% of the sequences marked as slippage by the arithmetic mean, geometric mean, and echo coverage methods, respectively. Thus, approximately 31% of the sequences marked as slipped by one strategy are not marked as slipped by the other.

The subsequence strategy can detect the highest score sequences of the suffix strategy (Figure 7). The first 2000 sequences pointed out by the suffix strategy were also pointed out by the subsequence strategy. However, this does not occur in the inverse direction. This result was expected: the highest score suffixes can be detected by the subsequence strategy, but the highest score subsequences are often lost by the suffix strategy.

The number of BLAST hits in the sequence set marked as slipped by the suffix strategy was greater than in sequence set of the subsequence strategy (Table 1). Comparing the methods, the arithmetic mean method shows more hits. The SUCEST method has a lower number of hits. Approximately 20% of the hits found in the complete sequences with no masking are due to vector hits in the sequence processed by our methods, and approximately 50% were for the sequences processed by the SUCEST method.

We can see (Table 1) that the impact of slippage removal in gene detection can be very small when the sequences are processed with our proposed methods. The percentage loss shown by them is no greater than 5%, while for the SUCEST method it is almost 90%. The reason for this difference is the discarding of criteria by their method. It should be recalled that only sequences that do not have poly-T tail, but that have poly-A tail of a given length, were preserved in further analyses.

The percentage loss of the suffix strategy is smaller than that with the subsequence strategy, but their values are very close. The geometric mean and echo coverage methods are closer in this aspect and both are better than the arithmetic mean method.

We conclude that the subsequence strategy is the best for the purpose of slippage detection. This strategy is more flexible and its effect on gene detection is very similar to the suffix strategy. However, we note that its complexity is quadratic, while the complexity of suffix strategy is linear.

Since the subsequence strategy shows the best results, we decided to perform further experiments with it. Thus, we defined the threshold values 1.90, 1.60 and 0.25 for the arithmetic mean, geometric mean, and echo coverage methods, respectively. These values are more restrictive and they reduce the size of the set of slipped sequences by nearly 15%. During the definition of these values, we confirmed our initial impression; the arithmetic mean method is the easiest to calibrate and the geometric mean method is the most difficult.

The choice of more restrictive values was motivated by the characteristics of our trimming strategy. We detected all artifacts independently and overlaps among them were not a problem. This generates sequence fragmentation, but only the longest sequence is preserved at the end of the process. We believe that the combination of detected artifacts can produce a better trimming, and that the more restrictive value will reduce the generation of false-positives.

The echo coverage method was elected as the best method. It appears to be capable of delimiting slipped regions with more precision than the other methods.

The geometric mean method, as mentioned previously, calculates scores that are very close and, as a result, its calibration is very difficult.

With the echo coverage method being considered the best, the arithmetic mean method must be analyzed more closely. Perhaps, in this type of method, the use of small values for the *minimum_echo_size* parameter can produce better results. We need to perform extra tests to evaluate the potential of this method under these conditions.

Moreover, we plan to work with the parameter *minimum_number_of_echoes* varying it for all proposed methods. We worked with the value eight for comparison purposes, but it does not mean that this is the best value. More tests should be carried out to evaluate this parameter.

We intend to carry on further tests with sequences of other organisms to determine whether the threshold values defined in this study apply to any organism, and to improve and validate the methods developed.

ACKNOWLEDGMENTS

The present study was developed at Scylla Bioinformatics (www.scylla.com.br) and was partially supported by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo - www.fapesp.br), grant numbers 2003/07748-9 and 2004/09417-2. We thank João Meidanis for correcting the manuscript.

REFERENCES

Adams MD, Kelley JM, Gocayne JD, Dubnick M, et al. (1991). Complementary DNA sequencing: expressed sequence tags and Human Genome Project. *Science* 252: 1651-1656.

Altschul SF, Madden TL, Schaffer AA, Zhang J, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25: 3389-3402.

Applied Biosystems (1998). Automated DNA sequencing - Chemistry guide. Part number: 4305080B.

Baudet C and Dias Z (2005). New EST trimming strategy. In: Lecture notes on bioinformatics (Setubal JC and Verjovski-Almeida S, eds.). Spring-Verlag Berlin, Heidelberg, West Germany. Brazilian Symposium on Bioinformatics (BSB 2005) 3594: 206-209.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*. 31: 365-370.

CPAN (2005). CPAN - Comprehensive Pearl Archive Network. http://www.cpan.org. Accessed October 2005.

Green P (2005). Phrap homepage: phred, phrap, consed, swat, cross match and Repeat-Masker documentation. http://www.phrap.org. Accessed October 2005.

Telles GP and da Silva FR (2001). Trimming and clustering sugarcane ESTs. *Genet. Mol. Biol*. 24: 17-23.

Vettore AL, da Silva FR, Kemper EL and Arruda P (2001). The libraries that made SUCEST. *Genet. Mol. Biol*. 24: 1-7.