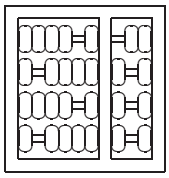

INSTITUTO DE COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DE CAMPINAS

Analysis of slipped sequences in ESTs Projects

Christian Baudet *Zanoni Dias*

Technical Report - IC-05-029 - Relatório Técnico

November - 2005 - Novembro

Analysis of slipped sequences in ESTs Projects

Christian Baudet * Zanoni Dias †

Abstract

Slippage is an important sequencing problem that can occur in EST projects. However, there are very few studies about it. In this work 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 291689 EST sequences produced in the SUCEST project [9], we performed comparative tests between the proposed methods and Telles and Silva Method [8]. 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. Comparing with the Telles and Silva Method, the advantage of our methods is that they do not discard the majority of the sequences marked as slippage, but, instead of it, only remove the slipped artifact from the sequence. The tests indicate that the "Echo Coverage Method" with *subsequence* strategy has the best compromise between slippage detection and calibration easiness.

1 Introduction

The objective of 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 (*Expressed Sequence Tag* [1]) is a cDNA (complementary DNA), that is a copy of an mRNA (messenger RNA) molecule. By sequencing a cDNA, we obtain a nucleotide sequence belonging to a gene that exists in the genome and was 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 cheaper than the other ones.

The chromatograms produced by those sequencing machines are processed by base-calling softwares that determine the base sequence of the EST. These softwares, usually, produce 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.

*Institute of Computing, University of Campinas, 13081-970 Campinas, SP.
†Institute of Computing, University of Campinas, 13081-970 Campinas, SP.

Slippage is an artifact type that can be found in EST sequences. Caused by sequencing process problems, slippage is a region that presents 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 to stay paired during the polymerization reaction and this can generate fragments with homopolymer regions of different length that have the same sequence after the region [3].

Although echoed bases sometimes appear with a high background noise, signal peaks are so high that base-calling softwares assign high quality values for bases that do not exist. This phenomenon prevents the removal of these regions by trimming methods based on quality.

The main objective of our work is to develop a set of trimming procedures for EST sequencing projects. Initial results of our research have been presented previously [4].

During our research, we observed that Telles and Silva were the only researches that carried a study on slipped sequences [8]. Their method defines an echoed region as a set of at least 5 identical consecutive bases. The product of echoed region lengths is evaluated for each sequence. If the echoed region length is equal or greater than 10, it contributes just 10 to the product. Sequences with product greater than 10^8 and echoed regions covering more than 20% of sequence length are considered slipped.

Once a sequence is considered slipped, an additional step, that 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 is 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 length lower 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 producing slippage detection improvement, we will discuss in this work three new alternatives, and compare them with the existing method.

2 Material and Methods

The three slippage detection methods proposed in this work 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 greatest suffix that reaches the threshold value. This strategy, that 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 begin and end positions clearly defined and that it is possible to discover them. Thus, the remaining sequence, that is not slipped, can be used in other analyses.

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

minimum_echo_size defines the minimum length that a group must have to be considered as a 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 highlighted is that all methods consider as valid echoed groups those that are only composed by 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 point to slippage artifacts that do not exist.

2.1 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 *suffix* strategy and parameters $minimum\_echo\_size = 4$ and $minimum\_number\_of\_echoes = 3$, 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

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

is identified as the best subsequence.

2.2 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 suffixes scores calculated for the same sequence above are

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

2.3 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 a 1 and every normal group is replaced by 0. The transformation of our example sequence is

000011010010.

After transforming, the method calculates the ratio between the number of 1s and the transformed region length. Then, the suffixes of our example have 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	000011010010	$4/12 = 0.33$

To perform tests with these methods, we employed the same data set used by Telles and Silva. This data set is composed by 291689 sugarcane EST sequences from the SUCEST project [9]. The average sequence length is 829.44 ± 182.60 bases. The average number of bases with PHRED (version 0.980904.e) [6] quality greater than 20 is 399.53 ± 182.60 bases. The sequences were produced from 26 libraries. The majority of the sequences (259325) were sequenced from the 5' end, while the remaining sequences (32364) were sequenced from the 3' end. We also implemented the Telles and Silva Method for comparison purposes.

In our tests we performed BLAST (version 2.2.11) [2] of slipped sequences against the Swiss-Prot database (release 46.6 - April 26, 2005) [5] to evaluate the influence of slippage removal in the gene detection process. Each slipped sequence was compared against the Swiss-Prot database in three different manners:

1. Complete sequence with no masking;
2. Complete sequence with vector masking, to measure the approximate number of hits found in the previous manner due to vector fragments;
3. Greatest contiguous subsequence that was not masked as slippage or as vector. Sequences with length lower than 100 bases were discarded.

The vector masking of the sequences was performed through the execution of cross_match (version 0.990319) [6] using the parameters `-minmatch 12` and `-minscore 20`.

The slippage masking for our proposed methods masked the longest slipped sequence that has a score greater than or equal to the threshold score that was chosen for the method. The slippage masking for the Telles and Silva Method was performed as described in their work.

For each pair method/strategy and Telles and Silva Method, we observed the percentage of sequences with at least one hit with e-value lower than or equal to 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) [7].

3 Results

Identifying suffix or subsequence with greatest score does not mean, necessarily, to find the whole slippage. Depending on echoed group distribution, the slippage score can be smaller than the score of a subsequence that is contained in it. So, the proposed methods need the definition of threshold values. The slipped region would be the greatest suffix or subsequence that has score greater or equal 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 $\{1, 2, \ldots, 10\}$. The parameter $minimum\_number\_of\_echoes$ was set to 8 for comparison purposes because this is the minimum number of echoed groups that are necessary to reach 10^8 in the Telles and Silva 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 split in 100-sequence intervals. Then the mean score of each interval was 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.

In order to illustrate better the behavior from distinct pairs method/strategy, we selected, for each one, a different base value as its slippage score threshold. We counted the number of sequences with score 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%, 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 of the Figure 4.

The second step of our tests was the comparison between the method results and the sequences reported as slipped by Telles and Silva method. This test was performed to evaluate the detection capacity through the contrast of different method results.

We implemented Telles and Silva Method (Method 4) according to the description found in their work. This implementation was used to process the same data set and 7213 sequences were marked as slippage. Notice that the processing was made with raw sequences and not with partially trimmed sequences as in their work.

The comparison was executed 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 Silva and

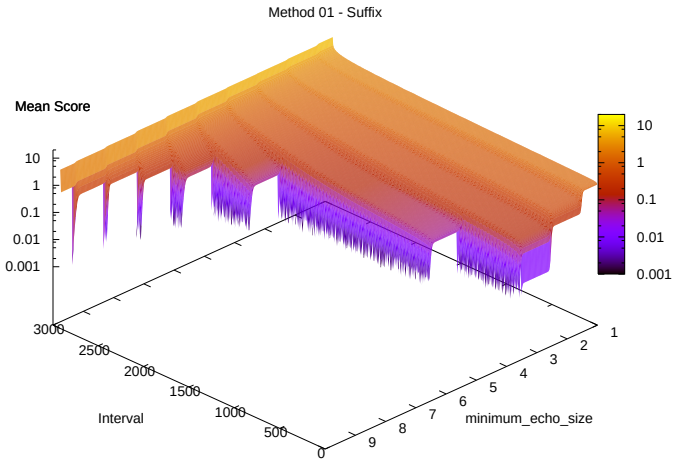

Figure 1: Arithmetic Mean Method executed with $minimum\_number\_of\_echoes = 8$, $minimum\_echo\_size = [1, 10]$, and *suffix* strategy. The results of each execution were sorted in ascending order and split in 100-sequence intervals, then their mean score are calculated. This graph shows the behavior of these intervals in each execution. Note that the mean score axis is in logarithmic scale.

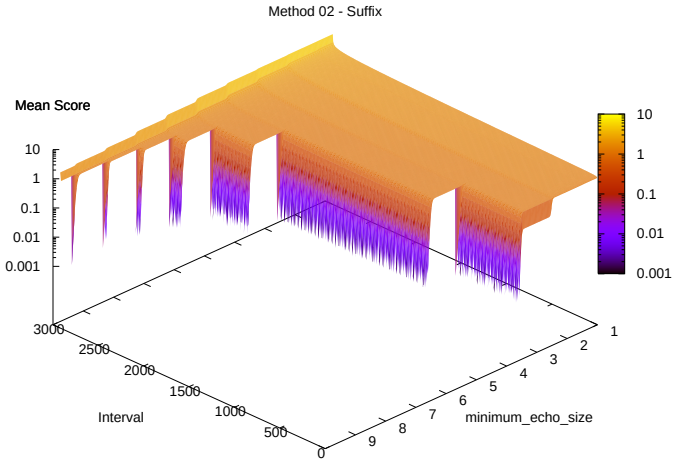

Figure 2: Geometric Mean Method executed with $minimum\_number\_of\_echoes = 8$, $minimum\_echo\_size = [1, 10]$, and *suffix* strategy. The results of each execution were sorted in ascending order and split in 100-sequence intervals, then their mean score are calculated. This graph shows the behavior of these intervals in each execution. Note that the mean score axis is in logarithmic scale.

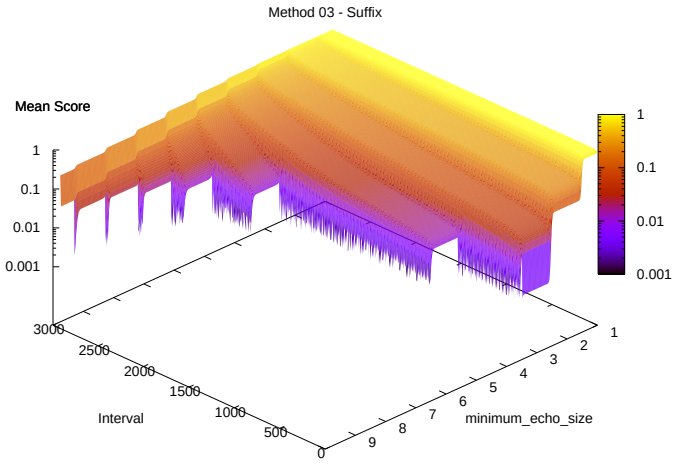

Figure 3: Echo Coverage Method executed with $minimum\_number\_of\_echoes = 8$, $minimum\_echo\_size = [1, 10]$, and *suffix* strategy. The results of each execution were sorted in ascending order and split in 100-sequence intervals, then their mean score are calculated. This graph shows the behavior of these intervals in each execution. Note that the mean score axis is in logarithmic scale.

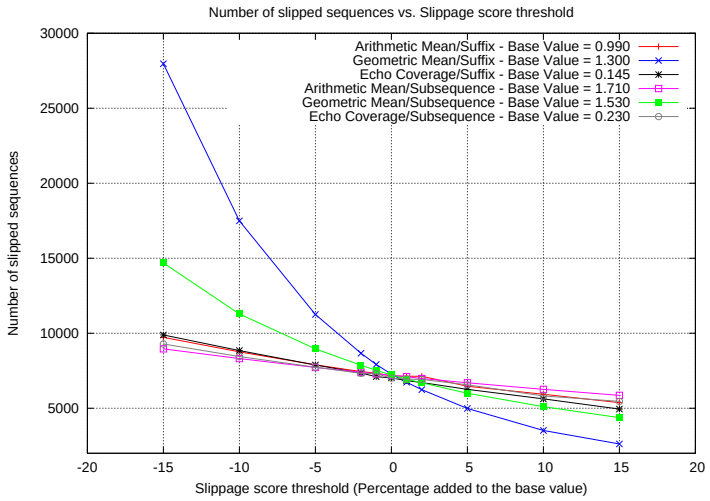

Figure 4: Number of sequences marked as slipped for each pair method/strategy using $minimum\_echo\_size = 5$. A base value was selected for each pair as the slippage score threshold. The threshold was varied by adding -15%, -10%, -5% -2%, -1%, 1%, 2%, 5%, 10%, 15% in the original base value.

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

For each pair method/strategy, we selected the 7213 sequence with largest score. The Venn-Euler diagrams shown by Figures 5 and 6 were constructed. They show the intersection of the sequence sets build by each proposed method and Telles and Silva Method, respectively, for *suffix* and *subsequence* strategies.

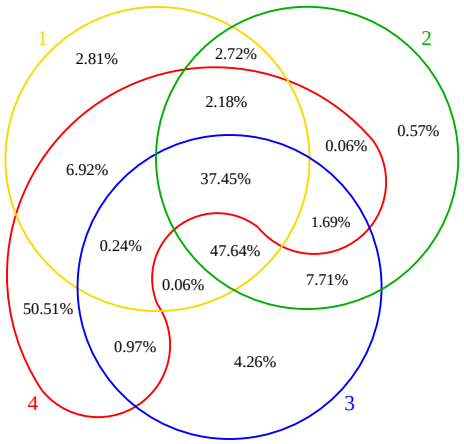

Figure 5: Venn-Euler diagram showing intersections of Arithmetic Mean Method (1), Geometric Mean Method (2), and Echo Coverage Method (3) sets using *suffix* strategy with $minimum\_echo\_size = 5$ and Telles and Silva Method (4) set. Each percentage value indicates the percentage of sequences of the method set that are inside of 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 compound by 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 to the Method 4 for the same region in the diagram.

The lists of 7213 sequences with greatest scores were also used in the third step of our tests. In this step we made the comparison between the strategy pair of each method. The intersection set of each pair has size 4976, 4969 and 4922, respectively, for Arithmetic Mean, Geometric Mean and Echo Coverage methods.

We sorted each list of each pair in descending order by score. We split them in intervals of 200 sequences and for each interval we count the number of sequences in it that were not found in the sequence set of the other strategy. The graph shown in Figure 7 shows the comparison between the two strategies.

The last test that we performed was the BLAST, as explained in the previous section, with the 7213 sequences marked as slipped by each one of the proposed pairs method/strategy and by Telles and Silva Method. The objective of it was to verify the influence of the slippage removal in the gene detection. The Table 1 shows the results of the BLAST runs and the loss percentage of hits when we compare the results of the first and second manners and the results of the second and third manners.

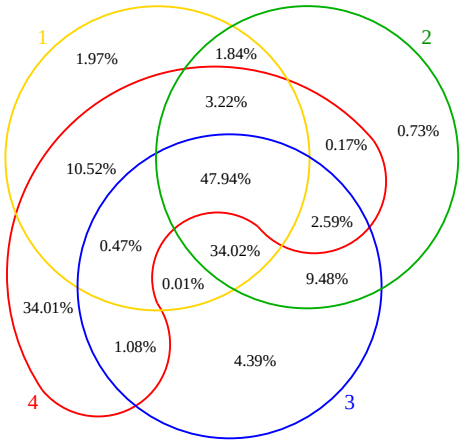

Figure 6: Venn-Euler diagram showing intersections of Arithmetic Mean Method (1), Geometric Mean Method (2), and Echo Coverage Method (3) sets using *subsequence* strategy with $minimum\_echo\_size = 5$ and Telles and Silva Method (4) set. Each percentage value indicates the percentage of sequences of the method set that are inside of 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 compound by 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 to the Method 2 for the same region in the diagram.

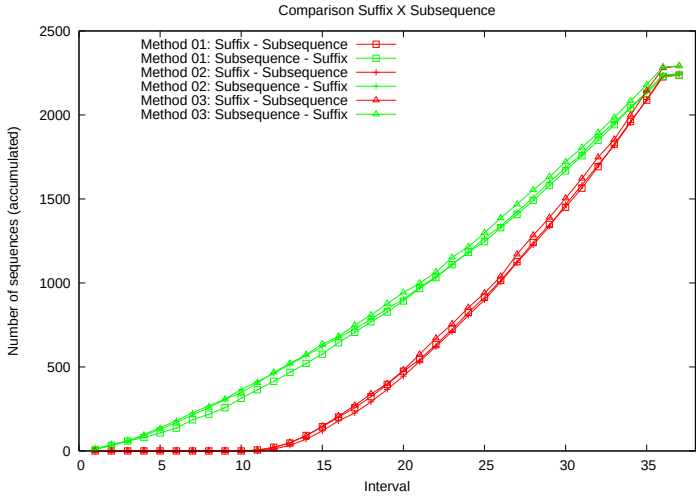

Figure 7: Comparison between the strategy pairs of each proposed method. Each result list was sorted in descending order. The firsts intervals have the sequences with greatest scores for each method/strategy. The green lines show the accumulated number of sequences that are into a *subsequence* set's interval and are not in the *suffix* strategy set. The red lines show the accumulated number of sequences that are into 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 pairs method/strategy (I - Arithmetic Mean, II - Geometric Mean, or III - Echo Coverage / *suffix* or *subsequence*) 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 - greatest contiguous subsequence with vector and slippage masking. The last two columns show the loss percentage of 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%

4 Discussion

Figures 1, 2, and 3 show the behavior of the score distribution of the *suffix* strategy, respectively, for Arithmetic Mean, Geometric Mean, and Echo Coverage methods. The score distribution of the *subsequence* strategy is similar despite the score increase caused by the shorter length of the slipped regions detected by it.

These graphs show that all methods have similar behavior as we change the value of the $minimum\_echo\_size$ parameter. The bands formed by intervals that have score mean zero 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 (8 in our tests) with length greater or equal than $minimum\_echo\_size$. If a sequence does not have a suffix that meets the criterion, it does not have a subsequence also, and vice versa. For example, if we get the results of Echo Coverage Method executed with $minimum\_echo\_size = 5$, we see that only 618 intervals have non-zero mean score.

Higher scores indicate more 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 small variation and evidences that this method will be hard to calibrate because small variations in the threshold value include or exclude a high number of sequences.

Analyzing the Figure 4 we can confirm this hypothesis. The Geometric Mean Method's curves grow quickly when we diminish the threshold value, while it does not happen for the other two methods. Looking for the Arithmetic Mean and Echo Coverage methods, we can see that their curves, in both strategies, are practically linear. Therefore, the Arithmetic

Mean Method is easier to calibrate because its curves have a lower inclination, i. e., the difference of the number of sequences between two threshold values is smaller than in the other methods.

Based on the surface graphs, we decided to set $minimum\_echo\_size = 5$ for all other tests. As we can see in these graphs, this is the first value that has the capacity of discard 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 great echoed groups.

For each one of the 6 pairs method/strategy, we took the 7213 sequences with largest score. This operation is equivalent to defining, for Arithmetic Mean, Geometric Mean, and Echo Coverage methods, respectively, the threshold values 0.9860, 1.3010 and 0.1429 (*suffix* strategy) and 1.7070, 1.5306 and 0.2286 (*subsequence* strategy).

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 happens 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 run with this strategy, they stay anchored in the same suffixes.

Another interesting fact that we see is that Geometric Mean Method is virtually covered by 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 Arithmetic Mean Method ($\sim$11%) or only by Echo Coverage Method ($\sim$5%).

Including the results of Telles and Silva 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%). Telles and Silva Method gives results closer to the *subsequence* strategy probably because, not being anchored to the sequence end, this strategy has more flexibility to find the echoed regions. Since Telles and Silva Method has the same characteristics, it is expected that it shows 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 graph of Figure 7 shows that the *subsequence* strategy can detect the highest score sequences of the *suffix* strategy. The first 2000 sequences pointed out by the *suffix* strategy were also pointed out by the *subsequence* strategy. However, this does not happen in the inverse direction. This result was expected: the largest score suffixes can be detected by the *subsequence* strategy, but the largest score subsequences are often lost by the *suffix* strategy.

The Table 1 shows that the number of BLAST hits in the sequence set marked as slipped by the *suffix* strategy is greater than in sequence set of the *subsequence* strategy.

Comparing the methods, the Arithmetic Mean Method presents more hits. The Telles and Silva method has the minor number of hits. Approximately 20% of the hits found in the complete sequences with no masking is due to vector hits in the sequence processed by our methods, and approximately 50% for the sequences processed by the Telles and Silva method.

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

The loss percentage of the *suffix* strategy is smaller than the presented by the *subsequence* strategy, but the value is very close. The Geometric Mean and Echo Coverage methods are closer in this aspect and both are better than the Arithmetic Mean Method.

The results produced in this work allow us to 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 must observe 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 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: Arithmetic Mean Method is the easiest to calibrate and Geometric Mean Method is the most difficult.

The choice of more restrictive values was motivated by the characteristics of our trimming strategy. We detect all artifacts independently and overlaps among them are not a problem. This generates sequence fragmentation but only the longest sequence is preserved in 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 generation of false positives.

Echo Coverage Method was elected as the best method. Its methodology appears to be capable of delimiting slipped regions with more precision than the other methods.

Geometric Mean Method, as mentioned previously, calculates scores that are very close and, as a result, its calibration is very hard.

Echo Coverage Method being considered the best, Arithmetic Mean Method must be analyzed with closer attention. Perhaps, in this type of methodology, 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_echo_size* varying it for all proposed methods. We worked with the value 8 for comparison purposes, but it does not mean that this is the best value. So, more tests must be carried out to evaluate this parameter.

We intend to carry on further tests with sequences of other organisms to verify whether the threshold values defined in this work apply to any organism, and to improve and validate

the methods developed.

5 Acknowledgments

This work 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, grant numbers 2003/07748-9 and 2004/09417-2. We wish to thank João Meidanis for manuscript revision.

References

[1] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, A. R. Kerlavage, W. R. McCombie, and J. C. Venter. Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project. *Science*, 252:1651–1656, June 1991.

[2] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[3] Applied Biosystems. *Automated DNA Sequencing - Chemistry Guide*, 1998. Part Number: 4305080B.

[4] C. Baudet and Z. Dias. New EST Trimming Strategy. In J.C. Setubal and S. Verjovski-Almeida, editors, *Lecture Notes on Bioninformatics*, volume 3594, pages 206–209. Springer-Verlag Berlin Heildelberg, July 2005. Brazilian Symposium on Bioinformatics (BSB 2005).

[5] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The Swiss-Prot protein knowledgebase and its supplement TrEMBL. *Nucleic Acids Research*, 31:365–370, 2003.

[6] P. Green. Phrap Homepage: phred, phrap, consed, swat, cross_match and RepeatMasker Documentation, March 2004. http://www.phrap.org.

[7] CPAN - Comprehensive Perl Archive Network, August 2005. http://www.cpan.org.

[8] G. P. Telles and F. R. da Silva. Trimming and clustering sugarcane ESTs. *Genetics and Molecular Biology*, 24(1-4):17–23, December 2001.

[9] A. L. Vettore, F. R. da Silva, E. L. Kemper, and P. Arruda. The libraries that made SUCEST. *Genetics and Molecular Biology*, 406:151–157, 2001.