INSTITUTO DE COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DE CAMPINAS

New EST trimming strategy

Christian Baudet *Zanoni Dias*

Technical Report - IC-05-09 - Relatório Técnico

May - 2005 - Maio

New EST trimming strategy

Christian Baudet [*] Zanoni Dias [†]

Abstract

Trimming procedures are an important part of the sequence analysis pipeline in an EST Sequencing Project. In general, trimming is done in several phases, each one detecting and removing some kind of undesirable artifact, such as low quality sequence, vectors or adapters sequence, and contamination. However, this strategy often results in a phase being unable to recognize its target because part of it was removed during a previous phase. To remedy this drawback, we propose a new strategy, where each phase detects but does not remove its target, leaving this decision to a post processing step occurring after all phases. Our tests show that this strategy can significantly improve the detection of artifacts.

1 Introduction

EST Sequencing Projects are developed with the objective of quickly obtain the gene index of an organism. The gene index is the list of the genes that exist in the genome of the organism.

An EST (*Expressed Sequence Tag* [1]) is a complementary DNA (cDNA), that is a copy of an mRNA molecule. When we sequence a cDNA we obtain the nucleotide sequence of 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. This last step is made through a single pass read in a sequencing machine and is one of the characteristics that make this technique cheaper than other existing techniques.

The chromatograms produced by sequencing machines are processed by base-calling softwares that determine the base sequence of the EST. These softwares, usually, produce quality values for each one of the bases determined. The quality value indicates the probability of error of the call. Normally, low quality bases are located in the sequence extremities. The error rate of base-calling programs can be as high as 3.0% [13].

One of the steps of EST production is the cDNA cloning phase. In this phase, vectors (plasmids, for example) and adapters are used for cDNA replication. After sequencing, we often find vector and adapter sequences in the nucleotide sequence obtained with the base-calling program.

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

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

The sequencing process is an automated process that is performed by specific machines. These machines determine base sequences by reading fluorescent signals and, sometimes, they cannot determine, with precision, the base value. A particular error happens when the machine finds a region that has a repetition of the same nucleotide with very high quality. In this situation the signals are so close and so strong that the machine cannot identify the signal peaks correctly and determines the existence of more than one base for a single position. This phenomenon in which the sequence seems "blurred" is called *slippage*.

Because of the characteristics pointed above, EST projects must submit the sequences to a sequence trimming process before analyzing them. The trimming process is a set of procedures that has the goal of removing from the sequence regions of low quality and subsequences that do not belong to the project target organism. This cleaning process must be performed because these subsequences can add errors to the data analysis. For example, a simple adapter sequence can determine if two sequences will be clustered together or not in a clustering process.

EST sequences suffer different kinds of contamination, depending, in part, of which of the many protocols was used in cDNA library construction. Thus, a method of verification of contamination must also be applied in the sequence processing phase.

Usually, each sequencing project executes its own sequence trimming and contamination detection protocols. Some projects perform a complete analysis, while others execute only low quality and vector trimming. This difference between the protocols compromises the comparison between sequences produced by different projects.

This work studies the difference between alternative methods of trimming and contamination detection, with the objective of creating a new set of procedures to improve the detection and removal of unwanted sequences.

2 EST Trimming

Trimming is the cleaning process of sequences produced by a sequencing project. It is responsible for the removal of regions that have low quality or are unwanted because they can cause errors in the sequence analysis phase. In this work, we denote these regions as artifacts.

The presence of artifacts in the sequence can have negative influence in the results of the analyzes of the data produced in the project. A low quality artifact, for example, represents a subsequence that has high error rates. It is not possible to guarantee that the nucleotide sequence of this artifact represents the real sequence found in the organism.

In a clustering process, overlapping is the most common criterion to cluster sequences. Vector, adapter, low quality and low complexity sequences can force the creation or separation of sequences through the erroneous sequence grouping caused by incorrect relationships added by these artifacts [14].

To avoid these problems, sequencing projects make use of different trimming techniques. Some projects make use of specific trimming softwares as ESTprep [12] or LUCY [5]. The latter is used by TIGR - The Institute of Genomic Research.

2.1 EST Trimming techniques

Each artifact type has appropriate detection and remotion procedures. Usually, after processing the sequence with these techniques, the remaining sequence has its length verified. If it is less than a minimum length value, the sequence is discarded and can not be used in sequence analysis processes like sequence clustering.

2.1.1 Low quality artifact trimming

Removal of low quality artifacts can be done in many different ways, which can be simple or more elaborated. The simplest solutions are usually fastest, an important factor when the data volume that must be processed is high. Thus, the decision on the strategy that will be used depends on the amount of time that can be spent with the task.

The quality value of a base determined by a base-calling software like phred [7] or TraceTuner [15] is based on the error probability of the base and is given by the equation $Q = -10 \times \log_{10}(error\ probability)$ [5]. Thus, a high error probability results in a low quality value.

A simple strategy to trimming low quality regions is the execution of an algorithm to determine a subsequence of maximum sum [9, page 106]. Each sequence processed by this algorithm has its low quality extremities removed. The software phred has a parameter to turn on this trimming strategy. The algorithm implemented by phred converts each quality value in an error probability value and tries to minimize the error probability sum of the sequence. Before executing the algorithm, each base has its error probability value decreased by 0.05. This value is equivalent to the quality value 13, the minimum value acceptable according to this implementation of the algorithm (phred version 0.020425.c).

Many projects perform the analysis through sliding windows. Usually, the sequence is covered base by base in both directions, from the extremities, with a window that has a fixed size. The window searches for regions that have at most a certain number of bases which have quality values lower than a minimum value. In the work developed by Telles and Silva [14], for example, the sliding window was 20 base long and the maximum number of bases with quality values lower than 10, allowed inside the window, was 12.

2.1.2 Vector and adapter trimming

One way to perform vector and adapter removal is using the software cross_match [7] to mask unwanted regions. This software receives as input the sequence that will be processed and the sequence of the target that must be identified. The regions that have an alignment with good score are masked with Xs. Thus, analyzing the X regions is possible to identify the vector and adapter artifacts. Telles and Silva make use of this technique and, additionally, classify the vector neighborhood into seven different classes.

More complex solutions, such as the implemented by LUCY, search in an adaptive way, guided by the quality of the analyzed region.

2.1.3 Poly-A and poly-T trimming

The trimming procedure developed by Telles and Silva performs the removal of poly-A/T tails after discarding vector and adapter regions. The poly-A/T tails are identified through the alignment between the sequence without vector or adapter and a long chain of As or Ts (typically 200 bases). A poly-A/T tail is identified if the alignment has a minimum score of 8 and a maximum distance of 10 bases from the sequence extremity. The alignment is made with the software swat [7] using the following scoring schema: 1 for a match, -2 for a mismatch and -8 for a gap.

2.1.4 Slippage trimming

As far as we know, Telles and Silva were the only ones to develop a procedure to detect slippage artifacts. This type of artifact was removed with the analysis of the alignments produced by swat. The alignments are analyzed for occurrence of abnormal repetition patterns.

3 Contamination detection

Sequence contamination is a serious problem in sequencing projects. Embarrassing occurrences have happened with frequency. For example, there are large-scale sequencing projects that had used libraries of clones highly contaminated and had to discard a huge amount of sequence. Other example happened in 1995, when the press announced the success of a research group in the extraction of DNA from dinosaur bones [17]. Nowadays, this announcement is seen as, at least, premature. The “dinosaur DNA” was compared to sequences of public databases and presented much more similarity to mammal sequences than to bird or crocodile sequences, suggesting that the DNA used in the analysis was a contamination with human DNA [18].

Many types of contamination are related with the protocols used in the production of the libraries and cloning [13]. Sequence contaminations can be separated in two groups: contaminations with DNA of other organisms and contamination with DNA that has origin in the same organism.

3.1 Contamination with DNA of other organisms

The vector used in the cloning process can be a source of contamination. Sequence rearrangements can insert vector sequence in the middle of the EST, producing a hybrid sequence. In this work, vector contamination is processed in the trimming phase.

A sequencing laboratory can conduct projects with different organisms. Accidentally, a clone library of one organism can be contaminated with sequences from other organism studieds in the same laboratory.

Material infected with virus or bacteria, or originated from organisms that have symbiotic relationships, can also produce contaminated sequences.

3.2 Contamination with DNA of the same organism

Sequences that belong to the target organism can also cause contamination. The contamination happens when there are ESTs containing subsequences that do not have origin in the processed mRNA.

Each EST is a sequence produced from some mRNA, but during the library construction process, the enzyme reverse transcriptase can capture rRNA to generate the cDNA. The same can happen with mitochondrial mRNA, chloroplast mRNA and premature mRNA. Genomic DNA can also be a contaminant.

3.3 Contamination detection techniques

Most projects use similarity to detect contaminated sequences [11]. Usually, the software BLAST [2] is used in this method. The sequence is compared against a contaminant sequence set. This set is composed by sequences of possible contaminants, rRNA, mitochondrial and chloroplast genes. The criterion of detection can vary among different projects.

Another technique is to use genome sequence characteristics to detect contamination. The strategy here is to classify the sequences into two groups (sequences that belong to the target organism and sequences that do not). This classification is made with the analysis of the values of the characteristics that the sequence has with the set of values, for the same characteristics, obtained from a set of sequences of the same organism and, optionally, from a set of contaminant sequences.

Many characteristics can be used. The work developed by White *et al.* [16] performs the detection with hexamers. Piazza and Setubal [11, 10] make use of a large variety of characteristics to improve the precision of the method.

4 Implemented trimming and contamination detection procedures

In this work we implemented two sets of procedures.

The first implemented set was based on our interpretation of the procedure described in the work of Telles and Silva [14].

The second set, called *New Schema*, was also based in the set proposed by Telles and Silva, but it has differences in sequence treatment. The major difference is in the input sequence of each step. In their set of procedures, all identified artifacts are removed before the sequence is submitted to the next step. In the new proposed set, the complete sequence is analyzed in each phase. The initial idea is to simplify detection methods for further refining of the techniques that did not show good results.

The motivation for this strategy was the observation of sequences processed by the procedures implemented by Telles and Silva. We observed that an artifact could be detected or not influenced by the detection or not of other artifact in a previous step. For example, if a vector is not identified, the detection of an adapter or a poly-A/T tail cannot occur because of the proximity criterion required by the method implemented.

Another point that we observed was the omission of artifacts that overlap with other artifacts. For example, certain adapters have an extremity that overlaps the extremity of the vector where they are inserted. In this case, when a vector is identified and discarded, the remaining adapter sequence is not detected because its size becomes too small for the identification criterion.

Figure 1 shows an example where the differences between strategies can be visualized. Sequence 1a) represents the one determined by the base-calling software. In this sequence we marked the position of all real artifacts that exist in it. Red color indicates low quality regions. Green color identifies vector regions and blue color pinpoints adapter regions. Yellow color marks the existence of a poly-A tail and gray color represents the insert that must be extracted from the raw sequence. Sequence 1b) shows the artifacts identified by the Telles and Silva procedure set. In this set, each artifact is identified after the removal of the artifact found in a previous step. Thus, the adapter and the poly-A tail were not detected because their remaining sequence dis not match with the criteria. So, the extracted insert goes to the analysis phase of the project with adapter and poly-A tail subsequences. Finally, sequence 1c) shows the artifacts determined by the new schema. The adapter artifact was determined correctly, but vector and the poly-A tail are not completely detected. It happens because they overlap low quality regions and the base-calling errors are more frequent in these regions. Nevertheless, the union of all artifacts permits the correct identification of the insert. This schema may not detect all artifact completely, however it is a good approach for estimating the artifact distribution in the sequence.

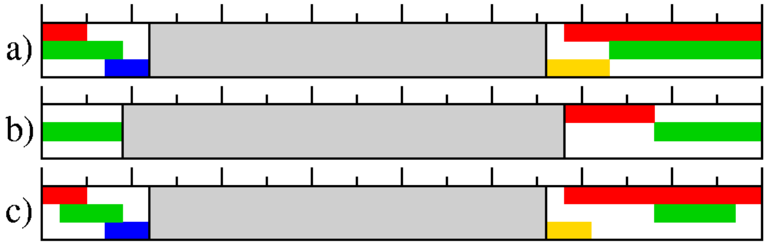

Figure 1: a) Artifacts found in a hypothetical EST sequence. Red color indicates low quality regions (top line), green color indicates the vector (middle line), blue color indicates the adapter (bottom line, left side) and yellow color indicates the poly-A tail (bottom line, right side). Gray color denotes the insert. b) Artifacts identified by the trimming procedure set of Telles and Silva. c) Artifacts identified by the New Schema.

ESTprep and Lucy perform trimming procedures that consider relationships between artifacts. Additionally, their trimming methods, in some steps, process only a subsequence of the original sequence because they remove a given type of artifact before looking for another.

The first step of the New Schema is ribosomal RNA detection. This is performed by BLASTing the sequence against a set of ribosomal sequences. If the sequence presents a hit with e-value lower than or equal to 10^{-10}, it is marked as contaminated and is discarded. This step is exactly the same as in the trimming set of Telles and Silva.

The second step is low quality trimming that is performed with a maximum subsequence algorithm similar to the one implemented in the software phred.

Vector detection is the next step. Detection is made with cross_match using, as parameters, 12 for minimum match and 20 for minimum score. Any subsequence identified is marked as vector. This strategy is different from the one adopted by Telles and Silva. It simplifies vector detection and obviates classification of the neighborhood of vector regions. In our method, the whole sequence is analyzed and this can result in more vector regions. This sequence fragmentation does not represent a problem because the last step of trimming preserves only the longest subsequence that does not contain an artifact.

After vector detection, the sequence is searched for adapters that were used in the sequence cloning process. Here, the criterion is to mark as adapter all regions that have an alignment with score greater than or equal to the size of the adapter minus four bases. The method implemented by Telles and Silva mask adapter sequences after the masking of vector sequences and remove them together in the vector trimming phase.

The following step is poly-A/T tail detection. The solution proposed is to perform an alignment of the target sequence with a long chain of 200 As or Ts using the software swat. All regions that have an alignment with a minimum score of 10 are marked as a poly-A or poly-T region. This strategy is also different in the Telles and Silva procedure set. Their solution performs the detection of tails in many phases (poly-T, poly-A, poly-T close to the left end, big poly-A, and poly-A close to the right end).

The last step of the trimming procedure set implemented for the New Schema is the identification of all maximum subsequences that do not contain any artifact and that have the minimum length of 100 bases. The subsequence must have at least 50 bases with quality greater than 20. Only the longest subsequence must be preserved. If there are several subsequences that meet the two criteria above, the method will choose the one with greater quality sum.

We are studying the detection of slipped sequences and we have not constructed a method yet. So, the New Schema does not implement this type of detection.

The contamination detection procedure is similar to the developed by Telles and Silva. First, vector regions are masked. After that, the sequence is compared against a set of contamination sequences with the software BLAST. If the sequence has a hit in a region that has a minimum length of 75 bases and an identity of 98.0% or higher, it is considered contaminated.

5 Testing data set

Tests are necessary for the validation of the methods developed in this work. We sought data sets that could be used in this validation process, and our preference was to use data from real EST sequencing projects.

Although there are many sequencing projects, only a few grant access to their sequence chromatograms. Usually, the projects publish the final processed data only.

In the site of the Cattle EST Project [4] we found available for download a set of chromatograms that were results of the sequencing of cDNA extracted from the placentas of *Bos taurus* individuals. There we also found the processed data (trimmed sequences and contigs of the clustering produced with the sequences).

Analyzing the chromatograms we found out that the placenta library was composed of 174 plates, each one with 96 wells. However, only 12620 sequences (75.55%) are available for download. These sequences are the ones that had not been discarded by the trimming process of the project. The trimming process implemented by the project discarded repetitive subsequences, low quality regions and ribosomal and mitochondrial sequences [3].

All chromatograms sequence that we obtained were submitted by the project in the dbEST [6]. Through the code identifier of each sequence, we could get from the NCBI (National Center for Biotechnology Information) the necessary information, like vector (pT7T3Pac) and adapter (5'-AATTCGGCACGAGG-3'), to perform the trimming process.

Neither the project site nor related articles give any information about contaminant organisms that could be found in the sequencing or the cloning laboratories.

The project clustered the sequences with the software CAP3, resulting in 8343 sequences in 2710 contigs. The 4277 remaining sequences were classified as singletons. In the site of the project we could also obtain the sequence and the composition of each contig.

6 Comparative analysis

The plate set that we had obtained was submitted to the two implemented procedures described in Section 4.

To perform the ribosomal contamination detection we constructed a repository of ribosomal sequences of mammals. The choice for mammal sequences is appropriate because *Bos taurus* is a mammal and ribosomal RNA is highly conserved among the species. Therefore, using sequences that are phylogenetically close, we improve the chances of detection. The repository was created with one sequence of the 28S subunit from *Mus musculus* (gi:53988) and one sequence of the 18S subunit from *Sus sucrofa* (gi:37956930).

Since all sequences had already been processed by the Cattle EST project, we expected that identification of rRNA sequences would not occur. However, the procedures implemented discarded 100 sequences. Of these discarded sequences, 98.0% had been identified as being of the subunit 28S and 2.0% of the subunit 18S.

Low quality trimming did not discard any sequence. This result was expected because of the observation that the trimming process performed by the Cattle EST project had discarded 24.45% of the sequences.

Both trimming procedure sets identified vector regions in 12461 sequences (99.53% of 12520 sequences preserved by the ribosomal detection phase).

The most dramatic difference between the two sets was shown by the adapter detection methods. The solution described by Telles and Silva found adapter regions in 91 sequences (0.7%), while our method detected them in 12311 sequences (98.3%). In their method, the vector is removed before the detection of the adapter, which, in this case, has a 6-base overlap with the vector sequence. Therefore, the remaining adapter could not be detected because its too short.

The size distribution of the adapters found by our method is shown in Table 1. Notice that the adapter length is 14 bases.

Table 1: Length distribution of the regions identified as adapter by the method implemented in the New Schema.

Region length	*Number of sequences*
10	3
11	12
12	17
13	24
14	12255
Total	*12311*

All 56 sequences that presented adapter region less than 14 bases were manually analyzed. This analysis evidenced that they were real adapter sequences. It was verified that their are close to the vector, but, because of low quality bases, their sequence could not be completely identified.

Our schema was also capable to detect more poly-A/T tails. Poly-A tails were found in 1957 (15.5%) sequences. Of these sequences, 1052 sequences (53.8%) had poly-A overlapping low quality regions. Of these, poly-A that had full overlap with low quality regions were found in 846 sequences (80.4%). The method implemented by Telles and Silva detected poly-A tails in 658 sequences (5.3%). Our method found poly-T tails in 955 sequences (7.6%) and the Telles and Silva method found in 718 sequences (5.7%).

We performed the detection of contamination comparing the sequences against the sequence of the organism *Escherichia coli*, a common contaminant in sequencing laboratories. We detected contamination in 14 sequences only. All contaminated sequences were previously discarded by the ribosomal trimming phase.

Table 2 shows the cluster size distribution of the clustering that were produced by the Cattle EST project with the software CAP3 [8]. It also shows the cluster size distribution of the clusterings that were created with the software phrap that received, as input, the sequences processed by the two implemented trimming procedure sets.

The comparison among the three clusterings could not be done with precision because of the difference of softwares. However, it was possible to get interesting results. There is a cluster of size 93 in the clustering of the project that grouped the ribosomal sequences discarded by the implemented trimming processes. However, we discarded 100 sequences and the cluster has 93 sequences. Of these remaining sequences, the software CAP3 classified 4 as singletons, grouped 1 sequence with a non ribosomal sequence and grouped 2 sequences with another non ribosomal sequence.

Comparing the clustering of the project and the clustering of the sequences processed with the Telles and Silva procedure set, we observed that 909 clusters were created in the latter one through the union or separation of clusters in the former. The same type of comparison was done with the clustering of sequences processed with the New Schema procedure set. It showed that 809 clusters were created through union or separation of

Table 2: Cluster size distribution of the clusterings Cattle EST Project (CAP3), Telles and Silva (phrap) and New Schema (phrap).

Size	*Cattle EST Project*	*Telles and Silva*	*New Schema*
1	4418	4644	4653
2	1384	1431	1419
3	590	601	590
4	254	237	239
5	130	119	116
6	69	66	65
7	41	39	43
8	40	39	36
9	13	11	9
10	9	9	9
11	6	6	7
12	2	5	5
13	7	6	6
14	2	4	3
15	3	2	3
16	1	2	2
17	4	2	3
18	1	0	0
19	1	0	0
20	2	1	2
21	1	1	0
23	0	1	2
24	1	0	1
25	2	0	0
27	1	0	0
29	1	0	0
30	1	0	0
33	1	1	1
45	0	1	1
48	1	0	0
93	1	0	0
Total	*6987*	*7228*	*7215*

clusters of the project clustering. This results evidences that the New Schema clustering is closer to the Cattle EST Project clustering.

When this comparison was made with the two phrap clusterings, we observed that 51 clusters were created in the New Schema clustering through the union or separation of clusters of the Telles and Silva clustering. Comparing their total number of clusters, we can see that the New Schema clustering privileged sequence grouping. This happens, mainly, because of the better adapter detection.

The New Schema did not perform slippage analysis. However, this fact did not impair our analysis. The Telles and Silva method found just one sequence that had its 3' end slipped. This sequence was classified as singleton by the three clusterings.

7 Conclusion

Our study evidences that the possibility of improvement in the trimming procedures is real. The New Schema proposed shows that the strategy of performing the detection of the artifacts individually, without constructing relationships among the different types of artifacts, can produce good results.

Some points are not covered by our procedures. The detection of slippage was not developed and we need to make more analysis to identify imperfections (false positive and false negative artifacts).

In this first version we did not developed an alternative for sequence contamination detection. One possibility, that we have in mind, is to vary the identity and covering criteria. Another possibility is to make use of sequence characteristics to complement the information about a supposed contamination.

The comparison among clusterings must be taken with care because of the difference in softwares used. We need to perform the clustering of the sequences processed by our method using CAP3, to make a better comparison. Also, a comparison with ESTprep and Lucy will be interisting.

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] M. R. Band, J. H. Larson, M. Rebeiz, C. A. Green, D. W. Heyen, J. Donovan, R. Windish, C. Steining, P. Mahyuddin, J. E. Womack, and H. A. Lewin. An Ordered Comparative Map of the Cattle and Human Genomes. *Genome Research*, 10:1359–1368, 2000.

[4] Cattle EST Project - The W. M. Keck Center for Comparative and Functional Genomics, University of Illinois at Urbana-Champaign, January 2005. http://titan.biotec.uiuc.edu/cattle/cattle_project.htm.

[5] H. Chou and M. H. Holmes. DNA sequence quality trimming and vector removal. *Bioinformatics*, 17:1093–1104, 2001.

[6] dbEST – The International Expressed Sequence Tags Database, July 2004. http://www.ncbi.nlm.nih.gov/dbEST.

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

[8] X. Huang and A. Madan. CAP3: a DNA sequence assembly program. *Genome Research*, 9:868–877, 1999.

[9] U. Manber. *Introduction to Algorithms*. Addison-Wesley, 1989.

[10] J P. Piazza. Uma metodologia para determinação do organismo de origem das seqüências de DNA com aplicação em projetos EST. Master's thesis, University of Campinas, Brazil, 2004. In Portuguese.

[11] J. P. Piazza and J. C. Setubal. New ways for automatic detection of contaminants in EST projects. In S. Lifschitz, editor, *Proceedings of Workshop of Bioinformatics (WOB'2003)*, Macaé - RJ, Brazil, December 2003.

[12] T. E. Scheetz, N. Trivedi, C. A. Roberts, T. Kucaba, B. Berger, N. L. Robinson, C. L. Birkett, A. J. Gavin, B. O'Leary, T. A. Braun, M. F. Bonaldo, H. P. Robinson, V. C. Sheffield, M. B. Soares, and T. L. Casavant. ESTprep: preprocessing cDNA sequence. *Bioinformatics*, 19(11):1318–1324, November 2003.

[13] R. Sorek and H. M. Safer. A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Research*, 31(3):1067–1074, 2003.

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

[15] TraceTuner. http://www.paracel.com/sas/tt.htm.

[16] O. White, T. Dunning, G. Sutton, M. Adams, J. C. Venter, and C. Fields. A quality control algorithm for DNA sequencing projects. *Nucleic Acids Research*, 21:3829–3838, 1993.

[17] S. R. Woodward, N. J. Weyand, and M. Bunnel. DNA sequences from cretaceous period bone fragments. *Science*, 266:1229–1232, 1994.

[18] H. Zischler, M. Hoss, O. Handt, A. von Haeseler, A. C. van der Kuyl, J. Goudsmit, and S. Paabo. Detecting dinosaur DNA. *Science*, 268:1191–1193, 1995.