REMOVING CONTAMINATION FROM GENOMIC SEQUENCES BASED ON VECTOR REFERENCE LIBRARIES

Caner Bagci, and Jens Allmer

Molecular Biology and Genetics, Izmir Institute of Technology
Gulbahce Campus, Urla, Izmir, Turkey
phone: + (90) 232 750 7517, email: jens@allmer.de
web: http://bioinformatics.iyte.edu.tr

ABSTRACT

DNA is often sequenced after being cloned into a vector since this provides the possibility for using standard primers and removes the need to develop custom primers. In this way a certain amount of vector is sequenced along with the sequence of interest. Unfortunately, occasionally these contaminating vector sequences find their way into public databases as part of submitted sequences. It has been pointed out that SeqClean, a program used to remove vector contamination from sequences, does not take into account that vectors are circular structures. A workaround has been presented before, but we were able to simplify the process and, additionally, we provide an implementation. We further applied our method to a test set of EST sequences and also analyzed the amount of contamination found in the EST sequences available on NCBI.

1. GENERAL INFORMATION

Sequencing of nucleotides is simple with the use of primers, short nucleotide chains with known sequence, complementary to a part of the sequence that shall be sequenced. Historically, sequences were first sub-cloned into vectors with known sequence so that the vector sequences could serve for the development of complementary primers. In order to keep copies of sequences, cloning is still in use today and large sequence libraries distributed over millions of vectors with different cloned sequences are available. Depending on where in the vector the sequence of interest was inserted, and on where the primers are in relation to the insert, the transcript which will be sequenced may contain some amount of vector/primer sequence. This is to be expected and should be taken care of by removing all parts of the final sequence that originate from primer/vector. New developments have made it possible to sequence without the need to sub-clone first. For example short adapters can be ligated to the sequence of interest and each adapter comes with a known primer that binds to the adapter sequence. Naturally, parts stemming from the adapter/primer need to be removed from the final sequence. That sequence contamination can pose problems has been noticed early on and one of the first successful programs dealing with sequence contamination is a combination of RAPID, PHAT and SPLAT . Several programs, including LUCY , LUCY2 , Figaro , SeqTrim , DeconSeq , TagCleaner , cross_match (http://www.phrap.org/phredphrapconsed.html), SeqClean (http://www.tigr.org/tdb/tgi/software), VecScreen (http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html), and a homology based method have since been developed to perform this task and they have been compared recently . SeqClean (http://www.tigr.org/tdb/tgi/software) is one of the most successful programs used for removing contamination from nucleotide sequences . Most programs, including SeqClean, rely on a library, containing all sequences that are possible contaminants. It has been shown that SeqClean is not able to remove contaminants if they span the linearization point (where a circular vector was linearized such that it can be stored in a sequence library) . For this problem a solution has been developed, but doesn't seem to be publicly available . Another solution, along the same lines, seems to be available at NCBI which simply appends the first 49 nucleotides of each sequence to the end of each sequence derived from a circular vector (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html#Pseudo). When we first approached the problem we only wanted to change the library (instead of the sequences to be tested) in order to make SeqClean understand the circularity of certain vectors. We chose to modify the library instead of the query sequences since the library more rarely changes than query sequences and changing the library is therefore more efficient. We then realized that NCBI did exactly that. Unfortunately, when using the NCBI library we were restricted to use the NCBI default settings of 49 appended nucleotides which would restrict us to SeqClean's default settings. Furthermore, we were surprised when we realized that in the complete UniVec database (when downloaded, from ftp://ftp.ncbi.nih.gov/pub/UniVec/), which should contain library sequences appended with the 49 nucleotides of the 5' part of the sequence appended as a copy to the 3' end of the same sequence, only 8 out of 4264 sequences were actually correctly appended. We then devised our own algorithm which provides all necessary flexibility to solve this problem. Our algorithm, called Library Processor, can check for instances where a copy of a 5' part has been appended to the 3' end and reverse them. It can be used to append sequences of any length from the 5' part of the sequence to its 3' end. This process is customizable such that, for instance, linear vectors and adapters can be excluded from the list of sequences to be processed either using a general filter or a specific id list. The algorithm is available on our website and can be downloaded or used via JAVA™ WebStart. We believe that this is important and that such software should be available not only to process UniVec but also to deal with

978-1-4673-0878-6/12/$31.00 ©2012 IEEE

HIBIT April 19-22, 2012 Cappadocia, Turkey

many custom vector databases which are not supported by NCBI. In the future, we will extend our system with a custom sequence cleaning module.

2. MATERIALS AND METHODS

2.1 Data

We created an artificial dataset which contains a number of problems which software that offers the ability to clean short sequence reads from vector contamination needs to be able to solve. Chen and colleagues assessed the extent of contamination in the NCBI EST database in 2007 by extracting every 600th EST sequence. We decided to do a reassessment and downloaded all EST sequences available on NCBI, selected every 600th sequence (26.12.2011). Another dataset contained all publicly available EST sequences from *Papaver somniferum*. All three datasets are available for download from our web site (http://bioinformatics.iyte.edu.tr/libraryprocessor). Many sequence cleaning tools depend on a library containing potential sequence contamination. We downloaded a widely used vector library, UniVec from NCBI (ftp://ftp.ncbi.nih.gov/pub/UniVec/, 23.12.2011). Although NCBI claims to have appended 49 nucleotides to all circular vectors, we were not able to confirm that. Therefore, we used three libraries, which are available for download from our web site.

1. The raw UniVec library as downloaded from NCBI (rawUV)
2. The UniVec library with the all appendices larger than 20 nucleotides removed (cleanUV)
3. The 2. Library with each sequence appended to itself (appUV)

2.1 Software

SeqClean was used since it has been assessed to be one of the best tools to clean sequences from vector contamination [9]. The latest version was downloaded from http://www.tigr.org/tdb/tgi/software. In addition to SeqClean we devised a number of in-house scripts.

1. to extract every 600th sequence from the full EST dataset, scripts to remove
2. to find and remove self-appendices from the sequences in the rawUV library
3. to extract sequences from the rawUV library which should not be re-linearized

Functions 2 and 3, which are of importance for anyone performing sequence analysis or assembly, were integrated into the Library Processor which is available on our web site either as a download or available for direct use via the JAVA™ web start function.

3. RESULTS AND DISCUSSION

Since NCBI claimed that they appended all circular vectors by a sequence of 49 nucleotides, we first wanted to confirm this and checked whether the 4264 sequences in rawUV show such appendices. We tested all possible appendix lengths with all sequences starting from 1 to half of the sequence length (which would mean that the sequence was appended to itself). We expected lengths close to one to show high appendix rates (due to chance) and expected an exponential decay of such occurrences with increasing length if no appendices were added to the library. Figure 1 shows that our expectation was true and that at the beginning of the graph, with increasing appendix length, the occurrence decreases exponentially (note that scale is logarithmic). The next observation we made was that only 8 sequences were appended with a 49 nucleotide long sequence from their own 5' part (Figure 1).

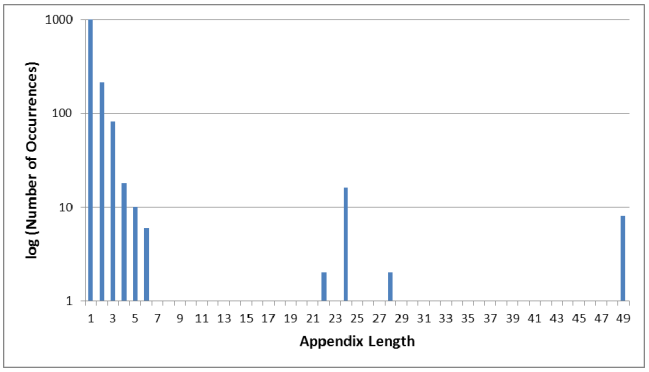

Figure 1: The appendices that were found in the rawUV library. Only the maximum appendix length is recorded for each sequence. The number of occurrences is presented in logarithmic scale. There were no occurrences of appendices larger than 49. Sequences that have no appendix (Appendix Length 0; 2903 occurrences) are not shown.

We then checked whether there were only 8 sequences in the rawUV library which derive from a circular vector. This was not the case which left us unable to explain why the majority of sequences which should have been appended, were not. A deeper analysis revealed that the vectors which contained appendices in the UniVec library were complete vectors. The eight sequences we refer to have the accession numbers J02400.1, M16192.1, L08860.1, J01749.1, V00604.2, M28829.1, X66730.1, and J02459.1 in rawUV. We examined a number of the vector sequence that were not appended and found out that these sequences were not full length, whereas full length sequences are available in nucleotide DB at NCBI (e.g.: GQ231553.1). Some vectors were further represented multiple times with different fragments. It is not clear to us why these partial sequences were stored in UniVec, while full length sequences are available. This leads to the same problem which Chen et al tried to overcome in 2007 and which we are tackling in this study. Another observation is an unexpected large peak at 24 nucleotide appendix length (Figure 1). A subsequent analysis exposed these sequences to be adapter sequences which were fully self-appended (e.g.: NGB00089.1). The same is true for the peaks at 22 and 28 nucleotides appendix length. Why this was done to these inherently linear sequences, by NCBI, however, we cannot explain. Interestingly this was not done for all adapter sequences (e.g.: NGB00018.1). It is,

however, clear, that this introduces new sequences which may lead to cleaning artifacts.
All appendices of length 20 or longer were removed from the rawUV library to create the cleanUV library which was used in the next step. The cleanUV library was also used to create the appUV library, by appending every sequence to itself, regardless of whether the sequence is from a circular or linear precursor.
These three libraries were used to compare the cleaning results for the three datasets, the artificial, the *Papaver somniferum*, and every 600th EST sequence from NCBI.
For each of these datasets SeqClean was used to remove vector contamination using all three libraries (Table 1).

Table 1: Differences in cleaning results for the three datasets used in this study versus the three vector libraries used in this study. The percentage of sequences cleaned and trashed is provided.

	rawUV	cleanUV	appUV
Every 600th EST	31.00	30.94	31.79
***P. somniferum* ESTs**	17.26	17.26	18.03
Artificial data	87.50	75.00	100.00

The amount of sequences cleaned/trashed, when seen in relation to the amount of sequences that are used in an assembly project, defines the dimension of the problem. Such contamination can form nuclei for clusters/contigs which in turn invalidate the overall assembly. That contaminated EST sequences lead to misassemblies and potentially faulty conclusions, has been pointed out . Table 1 shows that using the rawUV is more successful than using the cleanUV library (although the given precision does not reflect this for the *P. somniferum* dataset). In all cases cleaning with appUV is more successful than using any of the other libraries. Unfortunately, we have no means of establishing a ground truth since we do not know exactly which vectors have been used and whether they are actually available in UniVec. The results in Table 1 are based on a whole sequence view, either cleaned/trashed or not. A more precise measure could be the amount of nucleotides that was cleaned (Table 2).

Table 2: Differences in cleaning results for the three datasets used in this study versus the three vector libraries used in this study. The percentage of nucleotides cleaned and trashed is provided.

	rawUV	cleanUV	appUV
Every 600th EST	2.86	2.85	2.90
***P. somniferum* ESTs**	0.45	0.45	0.47
Artificial data	15.35	15.35	19.93

Table 2 shows again that using appUV for cleaning is the most successful among the three libraries. An interesting observation is that the *P. somniferum* dataset is significantly better when compared to the level of contamination in the general NCBI EST database. Naturally, the values in Table 2 are much smaller than in Table 1 since these records refer to nucleotide differences. Table 1 instead, measures on a per sequence entry basis. The above analysis was based completely on NCBI tools, but should be similar with DDBJ and EBI since sequences are shared on a daily basis. EBI also provides a vector sequence library (http://www.ebi.ac.uk/embl/Submission/vectors.html), but does not make any claim to add appendices. Our current understanding of EBI's vector library is that it contains full length sequences for commonly used vectors, as well as sequence fragments for new and/or partially sequenced vectors.

4. LIBRARY PROCESSOR SOFTWARE

The software we devised for library pre-processing was written in JAVA™ so that it may be used in any environment. We decided to add a graphical user interface since many of the potential users may not be familiar with scripting or the usage of a console application. The following steps are possible.

1. Removing existing appendices (given a user chosen minimum and maximum length for appendix detection)
2. Adding appendices to sequences in the library
    a. Possibility to selectively filter by keywords or id numbers
    b. Possibility to selectively filter by sequence length
3. Creation of multiple result libraries
    a. All appended sequences in one FASTA formatted file
    b. Sequences that were filtered in a FASTA formatted file.
    c. Sequences that were otherwise rejected from appending in a FASTA formatted file
    d. Combined file of cases a. – c.

Figure 2 shows the graphical user interface of the Library Processor software. We decided to keep the interface as simple as possible so that anyone can easily use it with the default settings, which should be sufficient in most cases.
In order to customize the library processing options, the settings (Figure 3) can be accessed by choosing Edit followed by Settings. Possible settings are divided into library cleaning options, sequence entry filtering options, and circularization options. Cleaning is important to remove prior appendices which would lead to artifacts if they remained within the sequence, unless they were full length appendices. To ensure that no appendices resulting from chance are removed, we use a default minimum value of 10 nucleotides which must be present at the beginning and end of the sequence to qualify as an appendix. In order to speed up search in large sequence libraries, a maximum appendix size can also be chosen (default is half the sequence length).

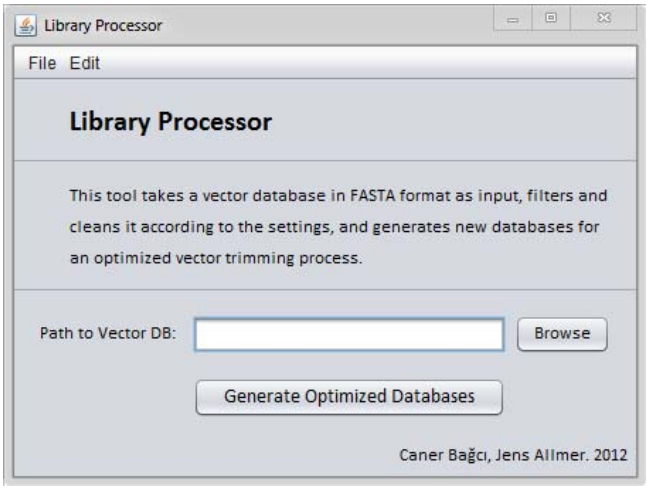

Figure 2: The main user interface of the Library Processor.

Once the library sequences have been pre-processed, some entries may need to be excluded from the virtual circularization. Two options are available, one filtering for keywords in the FASTA definition lines and one merely checking the length of the sequence in question.

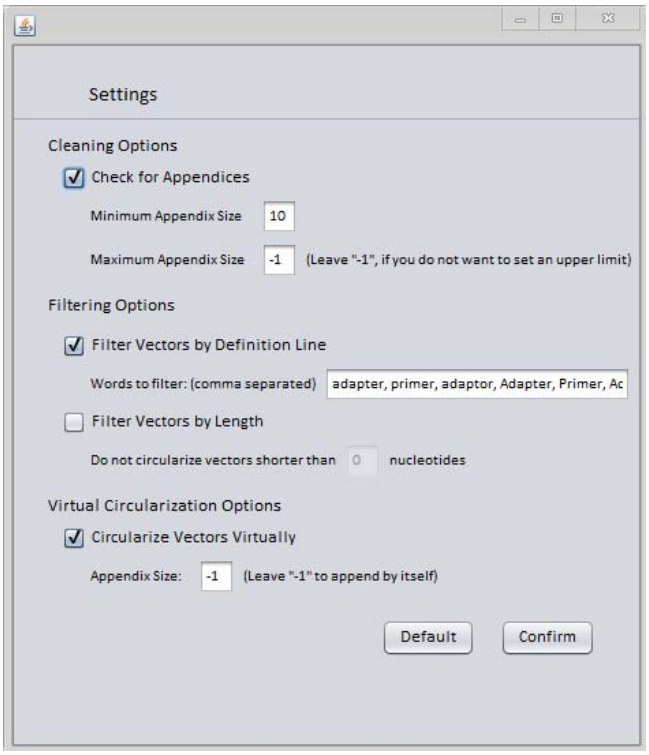

Figure 3: The settings panel of Library Processor which can be reached via Edit – Settings from the main panel (Figure 2).

The final step is the circularization of the sequences in the library which is only applied to sequences passing the filter criteria in the filtering options. Either the complete sequence is appended to itself or any custom appendix length can be chosen by the user. We believe it is essential to allow this feature since different studies may want to work with different cleaning settings for SeqClean which may call for other appendix lengths than the 49 nucleotides as imposed by NCBI.

5. CONCLUSION

We were able to show that UniVec has only 8 proper appendices, although 49 nucleotide appendices for all circular sequences are promised on the download page. We further proved that cleaning results improve with the quality of the vector library. We devised the Library Processor, a software, which can be used to pre-process any sequence library in FASTA format to re-linearize all or a selection of sequences. Unfortunately, several issues, which we consider problems, but which can as well be undocumented design choices, were uncovered when working with the UniVec library.

Our assessment of every 600th sequence from all available EST sequences from NCBI revealed that a large percentage is contaminated with vector sequences and that no improvement has been made since the 2007 assessment by Chen and colleagues [5]. Instead of pointing a finger at the submitter's data, we rather suggest that an automated function be used on NCBI which checks every incoming sequence for contamination and returns possibly contaminated sequences to the submitter with an appropriate report attached. Apparently, NCBI is thinking along the same line and in a recent paper they claim to be doing quality assessment of sequences submitted to GenBank . According to the paper contaminated sequences will enter the database but a quality report will be send to the submitters. Our analysis has shown that this implementation, at least existing since 1996 , has not worked and a new route has to be taken. We suggest rejecting any dubious sequence and forcing the submitters to review such sequences before resubmission. Another suggestion we can conclude from this study is that the problem of removing contamination from sequences is not yet closed which can be seen by two recent approaches by Robert and colleagues as well as Barker and colleagues . It thus seems to be essential that the removal of sequence contamination be formally evaluated and compared among methods before being integrated into an automated system dealing with sequence submissions.

6. OUTLOOK

There seems to be a desperate need to improve the UniVec library. We will, in the future, extract all known vector and adapter sequences to create a new sequence library which only consist of full length sequences. We theorize that this library will make a large change when used for cleaning of the currently available EST sequences. There is a need for a proper sequence library since most current sequence cleaning tools depend on one.

7. ACKNOWLEGEMENTS

We would like to thank Prof. Dr. Anne Frary for proof reading. This study was in part supported by the Turkish Academy of Sciences.

REFERENCES

[1] C. Miller, J. Gurd, and a Brass, "A RAPID algorithm for sequence database comparisons: application to the identification of vector contamination in the EMBL databases.," *Bioinformatics (Oxford, England)*, vol. 15, no. 2, pp. 111-21, Feb. 1999.

[2] H. H. Chou and M. H. Holmes, "DNA sequence quality trimming and vector removal," *Bioinformatics*, vol. 17, no. 12, pp. 1093–1104, 2001.

[3] S. Li and H.-H. Chou, "LUCY2: an interactive DNA sequence quality trimming and vector removal tool.," *Bioinformatics (Oxford, England)*, vol. 20, no. 16, pp. 2865-6, Nov. 2004.

[4] J. R. White, M. Roberts, J. a Yorke, and M. Pop, "Figaro: a novel statistical method for vector sequence removal.," *Bioinformatics (Oxford, England)*, vol. 24, no. 4, pp. 462-7, Feb. 2008.

[5] J. Falgueras, A. J. Lara, F. R. Cantó, G. Pé, and M. G. Claros, "SeqTrim – A Validation and Trimming Tool for All Purpose Sequence Reads," *Bioinformatics*, vol. 44, pp. 353-360, 2007.

[6] R. Schmieder and R. Edwards, "Fast identification and removal of sequence contamination from genomic and metagenomic datasets.," *PloS one*, vol. 6, no. 3, p. e17288, Jan. 2011.

[7] R. Schmieder, Y. W. Lim, F. Rohwer, and R. Edwards, "TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets," *BMC Bioinformatics*, vol. 11, p. 341, 2010.

[8] R. Sorek and H. M. Safer, "A novel algorithm for computational identification of contaminated EST libraries," *Nucleic Acids Research*, vol. 31, no. 3, pp. 1067–1074, Feb. 2003.

[9] Y.-A. Chen, C.-C. Lin, C.-D. Wang, H.-B. Wu, and P.-I. Hwang, "An optimized procedure greatly improves EST vector contamination removal," *BMC Genomics*, vol. 8, p. 416, 2007.

[10] D. a Benson, I. Karsch-Mizrachi, K. Clark, D. J. Lipman, J. Ostell, and E. W. Sayers, "GenBank.," *Nucleic acids research*, vol. 40, no. 2011, pp. 48-53, Dec. 2011.

[11] D. a Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "GenBank.," *Nucleic acids research*, vol. 24, no. Database issue, pp. 1-5, Jan. 1996.

[12] M. S. Barker et al., "EvoPipes.net: Bioinformatic Tools for Ecological and Evolutionary Genomics.," *Evolutionary bioinformatics online*, vol. 6, pp. 143-9, Jan. 2010.