Plant Tissue Cult. & Biotech. **17**(2): 161-172, 2007 (December)

PTC&B

Procedure to Identify and Submit cDNA Sequences to GenBank

George L. Britton Jr., Ahmad S. Islam, Yue Xuan and Kanagasabapathi Sathasivan*

Molecular, Cell and Developmental Biology, School of Biological Sciences, The University of Texas at Austin, USA-78712

Key words: Jute cDNA library, *Corchorus olitorius,* VecScreen, ORF Finder, Vector Contamination, GenBank, BankIt, Heat shock protein, TAIR WU-BLAST

Abstract

With rapid advances in DNA sequencing and large collection of such information available, it may be helpful to have a simplified procedure for analysis and submission of such sequence information. The process has been described in three steps: (a) data gathering and preparation, (b) data analysis, and (c) data submission to Genbank. The use of NCBI-tools such as *VecScreen* and *ORF finder* has been shown to be effective in detecting the vector contaminated sequences and in identifying the segments that code for protein with 'start-' and 'stop' codon, respectively. Another software called '*A Plasmid Editor*' is useful in reconstructing the complete DNA sequence of a gene from the sequences of complementary strand. The process of sequence data submission to the GenBank about the gene of interest is described in an easy-to-follow manner, including the final step of receiving an accession number from NCBI.

Introduction

The information technology boom has not only revolutionized the way people do business, but its influences have markedly changed the way biologists conduct their research. What was once laborious is now painless, because of the rise of the International Nucleotide Sequence Database Collaboration (INSDC) comprising three members: GenBank, European Molecular Biology Laboratory's European Bioinformatics Institute, and the DNA Data Bank of Japan. Their aim was to create an open access, annotated collection of all publicly available nucleotide sequences and their protein translations. Today, the three repositories have reached a significant milestone by collecting and disseminating 100 gigabases (10^9) of sequence data, and their combined database continues to grow at an exponential rate, doubling every ten months (Mehnert and Cravedi 2005).

Despite the rapidly growing wealth of information, problems arise with GenBank's DNA/RNA sequence submission web-based form, BankIt, because it

*Corresponding author. E-mail: sata@mail.utexas.edu

is not user friendly. GenBank provides excellent data mining tools, but their applications are not directly brought to the users' attention, resulting in unnecessary complexity. One such example led to the submission of several vector-contaminated inserts by one of the leading researchers (Sadhukhan et al. 2007).

Citing examples of cDNA sequences for *C. olitorius* var. O-4 (Wazni et al. 2007), this article aims to present the submission process in an easy-to-follow manner by describing the procedure and tools used to identify an unknown cDNA sequence for BankIt submission. To avoid any errors it is important to follow the general flowchart provided in Fig. 1. These points break down where to get information regarding the nucleotide sequence of interest, what information is important for submission, and how to use each program.

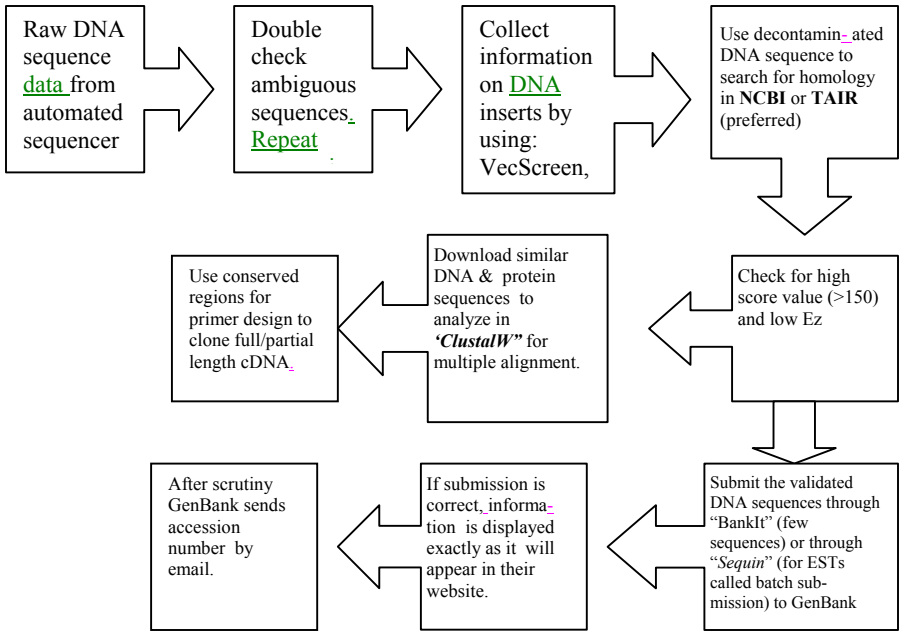


Fig. 1. Flow of DNA sequence analysis procedure.

DNA sequence information

Before submitting any information into BankIt, the researcher should retrieve all pertinent information regarding their unknown nucleotide sequence. Using a suitable example (Fig. 2A), this article describes the sequential steps on how to process an unknown base pair sequence with a view to annotate it prior to its submission into GenBank.

VecScreen - A diagnostic test: Before conducting any research regarding the nature of a cDNA sample it is imperative to check it for vector contamination by

using VecScreen provided by NCBI. Contamination is detected by running a BLAST sequence similarity search against the UniVec vector sequence database. The efficacy of this program is that any nucleotide sequence may be screened for vector contamination. Failing to recognize foreign inserts can lead to: 1) erroneous conclusions concerning the biological significance of the sequence, 2) wasting time and effort in analysis of contaminated sequence, 3) delaying the release of the sequence in a public database, 4) corrupt public databases with contaminated sequence (NCBI 2004).

VecScreen is not only helpful, but it is undeniably practical. To begin, simply enter the entire sequence in the provided box and click "Run VecScreen." One of two events will occur. Either "No significant similarity found" will result indicating there is no vector contamination, or a graphical display will appear showing the extent and distribution of vector contamination (Fig. 2b) along the sequence. In addition, the results will explicitly give the total size and relative position of the vector within the entire sequence. Once this is established, it is imperative to make the correct adjustments to the sequence of interest by removing the detected contaminated segments. VecScreen not only examines the quality of a given sequences of nucleotides, but it also shows the total amount of nucleotides within the insert which is useful information to enter into BankIt.

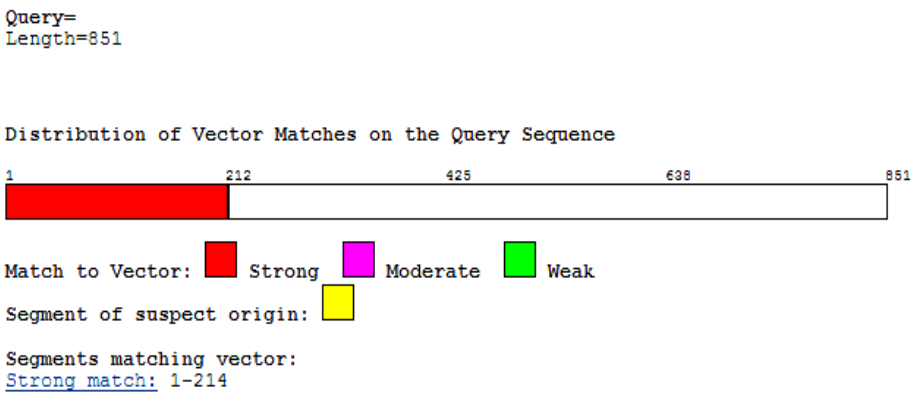
Open reading frame (ORF) finder: Once the vector sequence is removed, the insert sequence can be further analyzed using NCBI's ORF Finder (http://www.ncbi. nlm.nih.gov/projects/gorf/). This tool uses graphical analysis to find all open reading frames of a selectable maximum size in a given sequence (Tatusov and Tatusov 2007).

To run the program, enter the entire insert sequence into the provided box and click the box "ORFind." Next appearing on the screen will be sets of turquoise highlighted segments complemented with corresponding information regarding the reading frame (Fig. 3). BankIt will require each piece of information such as the reading frame, total length of coding sequence, and the positions of the start and the stop codon along the insert sequence. Additional information needed to complete the BankIt form will be the coding amino acid sequence, which can be found by clicking on the turquoise bar of interest. The reading frame consisting of only the amino acids can be entered in the Bankit form for completion.

A plasmid editor (ApE): ApE, or "A Plasmid Editor," is a useful application allowing DNA analysis to be a much less painful task. This simple and easy-to-use software is freely available online at http://www.biology.utah.edu/ jorgensen/wayned/ape/. For instance, the 5′ to 3′ reverse complement tool alone reduces analysis time considerably, as manual reversal of the complementarity of the sample would have been tedious and prone to human

AGTCGTATTACGGACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAAC
CCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAG
CTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGC
GCAGCCTGAATGGCGAATGGCGCTTCGCTTGGTAATAAAGCCCGCTTCGG
CGGGCTTTTTTTTGCAGAACATTGGTTTTTTTTTGGTTCTGATTTTCCAATGAGTAG
TAGTTGCATTAAACTTGATGTTCATACTGATGATCAAACTCCCCAGAAATGGTGCA
TTTCGTTAGCCGAAGACGTGTTCAAGAGATTTCTCTCTCAGGGTAATCCAACATTG
CATAAGGTATTTGGAGAAGGATCATTGTTTAGTCCCTTGTTGTTTGGGAAATATTTC
GATCCCTCAGACGCCTTTCCCCTGTGGGATTTTGAATCAGATAGCTTATTATCTAAT
CTAAGGAACTCTGGCAAAAGCACAGTTGATTGGTTTCAGACAGACCAAGCTTATG
TTCTTAAAGCAGAACTACCAGGATTAGGGAAAACTAATGTACAAATCCATGTGGA
AAAAGGGAAAATTGTGGAAATTAGTGGACAACTAAAGCAGCAAAGAGAATCCA
AGACAAAAGATTGGAGAAGCTGCAATTGGTGGGAATATGGATATGTAAGAAGGC
TTGAACTGCCTGAAGATGCAGATTGGAAAAGAATAGAGGCATACCTCAGTAATG
ATGTGCTTCTAGAGATTAGAATTCCCAGAAATCCTCTACATACTGATTTTCCTGAA
GCTGCAGTTGGCAAAATTTCAGAATGAATGTGAAAATCTGA

2A



2B

GCAGAACATTGGTTTTTTTTTGGTTCTGATTTTCCAATGAGTAGTAGTTGCATTAAA
CTTGATGTTCATACTGATGATCAAACTCCCCAGAAATGGTGCATTTCGTTAGCCGA
AGACGTGTTCAAGAGATTTCTCTCTCAGGGTAATCCAACATTGCATAAGGTATTTG
GAGAAGGATCATTGTTTAGTCCCTTGTTGTTTGGGAAATATTTCGATCCCTCAGAC
GCCTTTCCCCTGTGGGATTTTGAATCAGATAGCTTATTATCTAATCTAAGGAACTCT
GGCAAAAGCACAGTTGATTGGTTTCAGACAGACCAAGCTTATGTTCTTAAAGCAG
AACTACCAGGATTAGGGAAAACTAATGTACAAATCCATGTGGAAAAAGGGAAAA
TTGTGGAAATTAGTGGACAACTAAAGCAGCAAAGAGAATCCAAGACAAAAGATT
GGAGAAGCTGCAATTGGTGGGAATATGGATATGTAAGAAGGCTTGAACTGCCTGA
AGATGCAGATTGGAAAAGAATAGAGGCATACCTCAGTAATGATGTGCTTCTAGAG
ATTAGAATTCCCAGAAATCCTCTACATACTGATTTTCCTGAAGCTGCAGTTGGCAA
AATTTCAGAATGAATGTGA AAATCTGA

2C

Fig. 2A. Unknown nucleotide sequence with vector contamination (in bold). The user must enter the entire unknown sequence into VecScreen. 2B. Graphical display of vector contamination for unknown sequence. To fix the contamination problem simply excise the sequence showing a strong match (base pairs 1-214). 2C. Nucleotide sequence after removing the vector sequence.

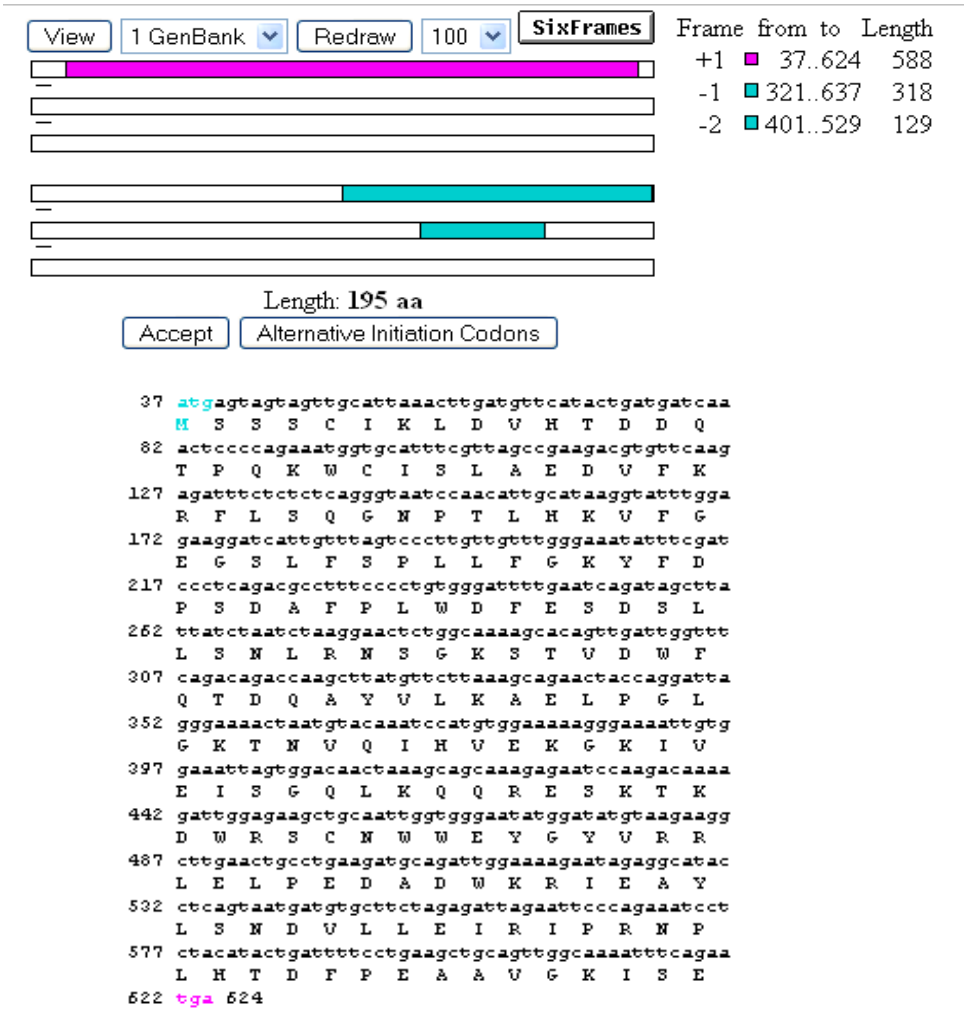


Fig. 3. OFR Finder results after entering a decontaminated nucleotide sequence. The highlighted pink area shows the reading frame (+1), where the coding sequence begins and ends along the sequence (from 37 to 624), the exact length of the coding sequence (588), and the translated amino acid sequence.

error. The other function allows for simulated restriction digests and displays predicted electrophoresis band patterns based on restriction sites found in the segment of DNA. However, only two functions were necessary for the purposes of our jute inserts - the aforementioned reverse complement tool and the DNA alignment tool.

Obtaining a quality sequence: Analysis typically starts with obtaining two relatively clean sequences obtained by using T3 and T7 or similar primers separately. As shown in Figure 4, the T3 (or forward) primer in our cDNA clones, provides sequences of the insert in the correct coding strand starting with an

EcoRI restriction site "GAATTC" towards its beginning and a poly-A tail towards its end as shown:

T3 Primer 5'----------EcoRI site-------cDNA Insert Sequence--------
AAAAAAAA---XhoI site----------T7 primer-3'

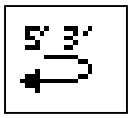
All of the sequences up to and including the Eco RI-restriction site as well as the sequences after the poly A tail are discarded, as they are not part of the insert and not of our interest.

However, if the insert is too long or if the sequencing quality is not ideal, often times it is found that the T3 sequence never reaches the poly A tail. Instead, the sequence might end with a segment saturated with many N's.

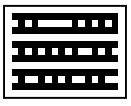
Yet, if there are only a few of these indeterminate bases, it is acceptable to search for these in the ICMB core facility-provided chromatogram and look for the closest matching base to fill it in. Many times, the uncertainty in the reading of the base is due to spill-over effects from the previous peak, which masks a peak of another color. If this is the case, then the N could be rather confidently determined to be the peak masked by the spill-over.

Luckily, even a T3 sequence that ends with many interspersed N's is not a problem if a good T7 sequence is available, in which case it is possible to fill in the missing part of the insert. This is largely because the T7 (or reverse) primer sequences from the opposite end of the strand complementary to the one sequenced by T3. This means the T7 sequence should start with a poly-T toward the beginning and, if it goes far enough, finish up with "GAATTC" toward the end. This way, the T7 sequence could provide information toward the end of the sequence in its correct orientation and the T3 sequence could provide information toward its beginning. But how do we stitch these two sequences together and clean up the raw data?

Here is where the ApE program becomes useful. First, the T3 and T7 sequences are saved as two separate ApE files, and the portions with many N's are cut away in either file. Then, the T7 sequence is treated with the reverse-complement tool (Fig. 4a) One way to check that it is in the correct orientation is to look for a poly-A tail in the treated sequence - you may need to re-sequence with T7 if the Poly-A is not there. Next, the DNA alignment tool is used (Fig. 4b).



4a: Reverse-complement tool.



4b: DNA alignment tool.

At this step, it is crucial to check that the T3 and the T7 sequences are being matched up against each other and not themselves. It is also helpful to check that the "reverse complement" box is not checked for either T3 or T7, as we have

already done that for the T7 sequence and the two sequences are now in the same orientation.

Next, clicking "ok" brings the browser to a new window that shows the alignment results. Notice that the alignment, with matches indicated by a vertical line, is executed directly; this means that A's are matched to A's, G's are matched to G's, etc. (Fig. 5). This is also the reason why the T7 sequence must be reverse-complemented either by using the button before the alignment or by checking the box during the execution of the alignment. Also notice that the alignment should mostly be correct in the middle of the sequences - this is the area of overlap. Lack of overlapping between the two sequences indicates that there is still missing information in the middle requiring the sample to be sent again for sequencing. However, in most cases the insert is short enough for the T3 and T7 sequences to fit past the middle, resulting in an overlap.

The next step uses one of these two sequences to construct the full sequence of the DNA insert at hand. At this point, it is easy to see that one strand (T3) has a good sequence before the overlapping region that the other lacks; the other sequence (T7) should have an unambiguous sequence beyond the overlap. If one sequence (either T3 or T7) has no N's or skipped nucleotides for the overlapping region, then it would be wise to use that sequence as a starting point. Otherwise, start with T3 or T7.

After copying the sequence of interest as the "skeleton" of the full sequence into a new window and saving it as a new ApE file, switch to the alignment results window and look for a short sequence of bases at the point where your "skeleton's" overlapping region begins or ends. For example, if you chose T3 as your "skeleton," then you should use the last 6 bases (at the end of the overlap). Flip over to the T7 window, which should already be in the correct orientation and with the poly-A, and do a search on the sequence with the last 6 bases just obtained from the T3 sequence. This should be at the end of the overlap. Next, copy everything after those six base pairs from the T7 sequence and append it to the end of the T3 skeleton (Fig. 4b) Thus, use of more than one copy of the overlapping region is avoided in the final sequence.

DNA sequence identification

Analysis regarding an unknown sequence may be performed using NCBI's Basic Local Alignment Search Tool (BLAST), which enables researchers to find regions of similarity between sequences of interest and those available in the database. BLAST analysis allows researchers to compare a query sequence with a library or database of sequences and calculates statistical significances between matches. BLAST proves to be advantageous to researchers because it addresses a fundamental problem of the algorithm's emphasis on speed over sensitivity. This

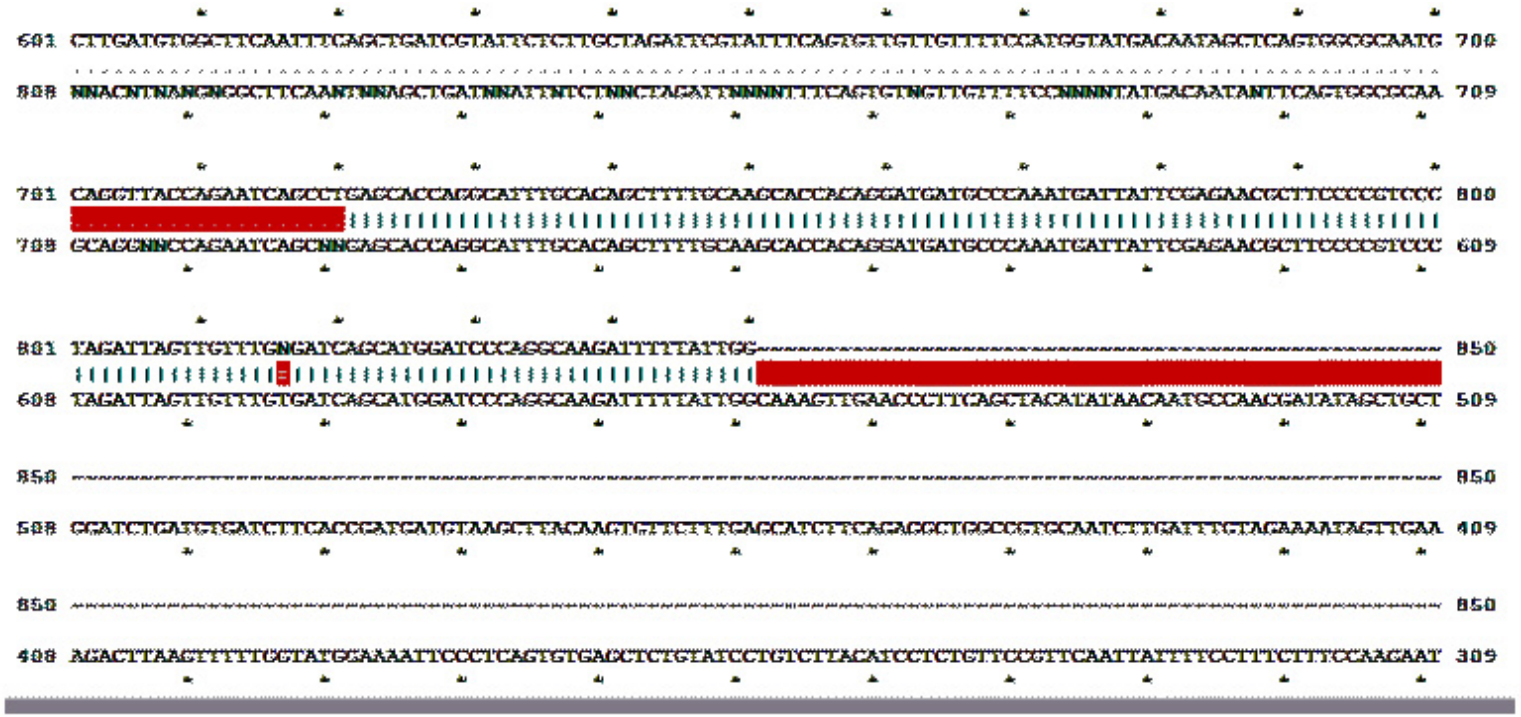


Fig. 5. The cDNA clone was constructed with Eco RI and XhoI in pBluescript vector will give the following sequences when sequenced with T3 and T7 primers. Alignment overlaps of T3 sequence with T7 sequence.

emphasis on speed makes the application practical, while searching massive genome databases.

Although NCBI developed the algorithm and provides the program on their website, its extensive database of sequences from a wide variety of organisms from various kingdoms and domains, may lead to an unfocused search. As a result, the searches may not provide meaningful homologies. As a solution, we found it effective to simply search in plant specific data bases such as TAIR (The Arabidopsis Information Resource, http://www.arabidopsis.org) using their WU-BLAST tool, *(http://www.arabidopsis.org/wublast/index2.jsp),* where the completely sequenced *Arabidoposis thaliana* genome is used for similarities between query sequence and that available in the database.

TAIR Wu-Blast: The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for the model plant *Arabidopsis thaliana*. It compares the query sequence with that of *Arabidopsis thaliana* to provide a clue as to what protein our nucleotide sequence most likely codes for.

To begin the analysis use the above provided URL that takes the user to the TAIR homepage. Next, move the cursor over the "Tools" menu bar and click on the link "WU - BLAST." First, enter the nucleotide sequence into the provided box entitled "Input." Now the user must decide what information they want by changing the upper drop down menu entitled "BLAST program." Refer to the WU-BLAST website http://www.arabidopsis.org/wublast/index2.jsp for all possible BLAST programs. Next, the user must decide what "Datasets" to use. While searching for nucleotide sequences, first try the option "AGI Transcripts (-introns, + UTR) (DNA)," and click "Run BLAST." In the event of "No hit," change the "Dataset" option to "Higher Plants EST (DNA)", while keeping the "BLAST Program" contents the same, and click on "Run BLAST." The resulting page will show a list of sequences producing high-scoring segment pairs (Fig. 6).

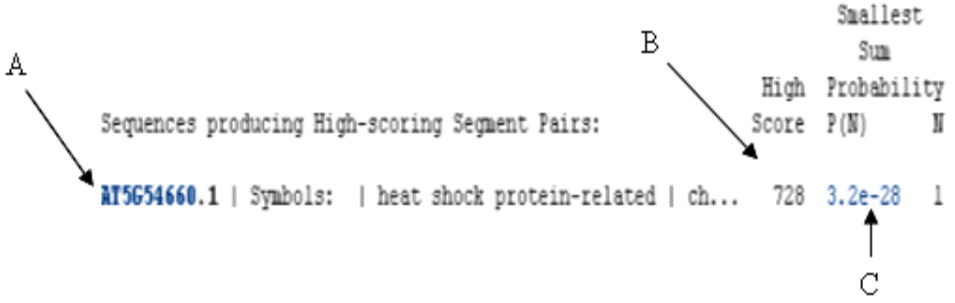


Fig. 6. Results from TAIR-WU BLAST show only one possible choice relating to the nucleotide sequence to a heat shock protein. Letter B points to the score value, while letter C points to the P-value. Click on the accession number (A) to inquire more information related to the heat shock protein.

The most likely match to the entered sequence will have the highest "High Score", while attaining the lowest "P-Value." While determining the best match one must bear in mind that the "High Score" should be over 150 with a "P-Value" no larger than 0.01. Once the user has determined the best fit applying the above criteria, record the corresponding values because these numbers will be useful for BankIt submission.

To retrieve further information for BankIt, click the link (shown as a representative gene model number, *AT*5G54660.1) that corresponds with the best-fit sequence segment. The page displays relevant information regarding what protein the nucleotide sequence translates. In this case it is the heat shock protein as indicated above.

DNA and protein sequence submission to GenBank: Once all data regarding the cDNA sequence are collected by using the above tools, the researcher may begin to submit their findings to GenBank. There are two separate routes for submission. One way is to use the downloadable program, Sequin, provided by NCBI, and is used for submitting multiple EST sequences in bulk. The second method uses an online web-based form called BankIt, which is used for submission of both full and partial cDNA sequences, one or a few at a time. This article discusses the BankIt format because sequence analysis tools are not directly provided to the user during the submission process.

To create a new submission form, visit the BankIt entry page at http://www.ncbi.nlm.nih.gov/BankIt/index.html. Enter the total length of the nucleotide sequence and press the icon “New.” Now the first page of the form presents itself. This half of the form is the easiest and most direct because it only needs contact-, reference-, source-, and DNA sequence information. Be sure to completely fill in all relevant boxes. Once the first page is completed, click the icon “Save This Form”. This will save the form to the hard drive enabling the user to utilize it for all subsequent BankIt submissions. Once saved, check the validity of the entered information by clicking on the icon “Validate and Continue.” Any errors will appear at the top of the BankIt page with links to where the problems are located. Once all warnings and errors are fixed, click the icon “Review and Submit” to view the second page of the form enabling the user to correct and modify the current submission. To do so, replace the number “0” with “1” in the topmost box while leaving all other boxes unchanged. Now press “Modify Submission” immediately below the boxes. Now, enter all relevant information gathered, when processing the DNA sample. While entering the amino acid sequence in the provided box, begin with methionine while removing the chain of amino acids before it. Once all data are entered and verified for accuracy, click the icon "Submit to GenBank." Immediately following the submission, GenBank will provide a BankIt number showing how the submitted

information will appear on their website. After scrutiny, GenBank sends an accession number of the submitted partial or full bp sequences of the gene of interest. Verify the information and make any changes or confirm the given information by email. This information can be updated later if necessary to make any additions or changes to existing data.

Conclusion

The many challenges of extracting useful information from the experimental data may be discouraging. However, using the tools mentioned, researchers may conduct well-grounded and thorough data analysis before submitting nucleotide sequences for publication to GenBank.

In summary, the process of DNA sequence analysis should be accomplished in three steps: 1) data gathering and preparation, 2) data analysis, and 3) data submission as illustrated in Figure 1 given in the form of a Flow Chart. When gathering and sorting data, one could use the above-mentioned tools in gleaning useful information from the data while discarding others. For instance, *VecScreen* allows for the detection of vector-contaminated sequences and the subsequent clean-up of the sequence; *ORF finder* helps identify the segment of the sequence that actually codes for protein. *A Plasmid Editor* aids in the further consolidation of useful information and the reconstruction of the complete cDNA sequence. Additionally, when examining the corrected data, one could take advantage of the various tools offered on the NCBI webpage as well as TAIR Wu-Blast for plant genomics and conduct a multi-dimensional analysis.

These methods make the process easier to create the individual profiles of new DNA sequences within the rich databases of NCBI, and this is the way that knowledge grows within the global scientific community.

Acknowledgment

Funding for the current work was partly provided by University CoOp Undergraduate Research Fellowship at The University of Texas at Austin. We are grateful to the GenBank Direct Submission staff, particularly Dr. Vincent Calhoun, who advised in the choice of software that yielded us the necessary information that was needed for submission of jute cDNA inserts. We are indebted to our undergraduate students, particularly, Mohamad W. Wazni, Matthew J. Talliaferro and Nabila Anwar for their help in the extraction of cDNA clones from *C. olitorius* and analysis of DNA sequences.

References

Mehnert R and **Cravedi K** (2007) Public Collections of DNA and RNA SequenceReach 100 Gigabases. National Library of Medicine. 22 Aug. 2005. National Insitute of Health. 26 Nov. 2007 http://www.nlm.nih.gov/ news/press_releases/ dna_rna_100_gig.html .

National Center for Biotechnology Information (2004) VecScreen." National Center for Biotechnology Information. 18 Oct. 2004. National Institute of Health. 10 June 2007 http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen. html .

Sadhukhan S, Basu A, Maity MK and **Sen SK** (2007) NCBI database for jute ESTs http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucest&cmd=search&term=Sen

Tatusov T and Tatusov R (2007). ORF Finder (Open Reading Frame Finder). National Center for Biotechnology Information. National Institute of Health. 9 June 2007 http://www.ncbi.nlm.nih.gov/projects/gorf/ .

Wazni, MW, Islam, AS, Taliaferro M, Anwar N and **Sathasivan K** (2007) Novel ESTs from a Jute (*Corchorus olitorius* L.) cDNA Library. Plant Tissue Cult. & Biotech. **17**(2): 173-182.

Websites Used in the Study:

A plasmid Editor (ApE) http://www.biology.utah.edu/jorgensen/wayned/ape/ ClustalW for Multiple Alignment: http://align.genome.jp/

National Center of Biotechnology Information http://www.ncbi.nlm.nih.gov/

Open Reading Frame (ORF) Finder: http://www.ncbi.nlm.nih.gov/gorf/ gorf.html

TAIR: The *Arabidopsis* Information Resources (TAIR: http://www.arabidopsis. org)

VecScreen. One of the mining tools available in the NCBI site useful for detecting the degree of vector contamination. http://www.ncbi.nlm.nih.gov/VecScreen/Vec Screen.html