APPLICATION NOTES

Genomics & Informatics Vol. 4(4) 170-172, December 2006

ESTin: A Program for Building a dbEST Submission File from Massive EST Sequences

Byungwook Lee[1,2]*

[1]Korean BioInformation Center, KRIBB, Guseong dong, Yuseong-gu, Daejeon 305-817, Korea, [2]Department of BioSystems, KAIST, Guseong dong, Yuseong-gu, Daejeon 305-701, Korea

Abstract

ESTin is an easy-to-use tool that allows EST researchers to build a dbEST submission file with EST sequences and related data. The advantages of ESTin are its user-friendly interface, various editing functions, and rigorous validation function.

Availability: ESTin is freely available at http://pat.kobic.re.kr/biodeposit/est/estin.htm

Supplementary information: A user manual is available on the ESTin website.

Keywords: dbEST, ESTin, EST submission

Introduction

Expressed sequence tag (EST) is generated by single-pass 5' or 3' DNA sequencing of clones randomly picked from cDNA libraries (Adams *et al.*, 1991). It represents a partial description of the transcribed portions of genomes, and thus can provide insight into transcribed genes in a variety of organisms (Boguski, 1995). Most EST sequences are deposited into dbEST, a central public EST repository, in order to obtain the deposition numbersfor publication and supply them to the bio-community (Boguski *et al.*, 1993). dbEST is the largest division of GenBank. Currently, it contains 38,933,423 entries that account for sixty-two percent of the entire GenBank entries (Benson *et al.*, 2006).

To submit EST sequences to dbEST, they should be converted into EST submission files, according to the dbEST format (http://www.ncbi.nlm.nih.gov/dbEST/how_to_submit.html), which can be sent to dbEST via email. There are four types of dbEST format: contact, library, publication, and EST. Each type has its own fields consisting of a capitalized field tag, followed by data. It is easy to make the contact, library, and publication files. However, it is a difficult job to build an EST file, becausemost EST projects produce large numbers of EST sequences. Although it is possible to prepare them with manual editing, this is a tedious task and can cause errors in the submission file. Therefore, in most cases, using programs is necessary for preparing EST submission files.

Although several programs have been developed for preparing EST submission files, most of them are integrated in large packages for EST processing and annotation such as ESTAP (Mao *et al.*, 2003), ESTWeb (Paquola *et al.*, 2003), and PartiGene (Parkinson *et al.*, 2004). Accordingly, to use this function, these packages and all the pre-required programs should be installed on the user's computer, which is an unnecessary job except for a package user. Moreover, most of these programs are capable of building EST submission files only with the EST sequences from their own package's output.

The author has developed ESTin, a program for building EST submission files to dbEST. The advantages of ESTin are its user-friendly interface, various editing functions, and a rigorous validating function.

Methods

ESTin was written in Microsoft Visual C++ 6.0 and runs on PC-compatible computers under the Microsoft Windows operating system. Its interface consists of a main window and several input dialogs. The user can input contact, library, publication, and EST data by clicking on the corresponding buttons in the main window. Figure 1 shows the main window and several input dialogs.

In preparing an EST submission file, the most significant aspect is the EST sequences. ESTin accepts FASTA-formatted EST sequences from external files and treats the first word in the description line of an EST sequence as the EST name.

Most ESTs have features that are included in the submission files. Some features are common to all ESTs, such as the Library name, whereas the others are unique to each EST such as row and column numbers. Users can assign common features to all ESTs or to a part of them during the importing step. To assign the unique features to ESTs, these features need to be converted to a tag=data form and added them to the

*Corresponding author: E-mail bulee@kribb.re.kr, Tel +82-42-879-8535, Fax +82-42-879-8519
Accepted 28 Nov 2006

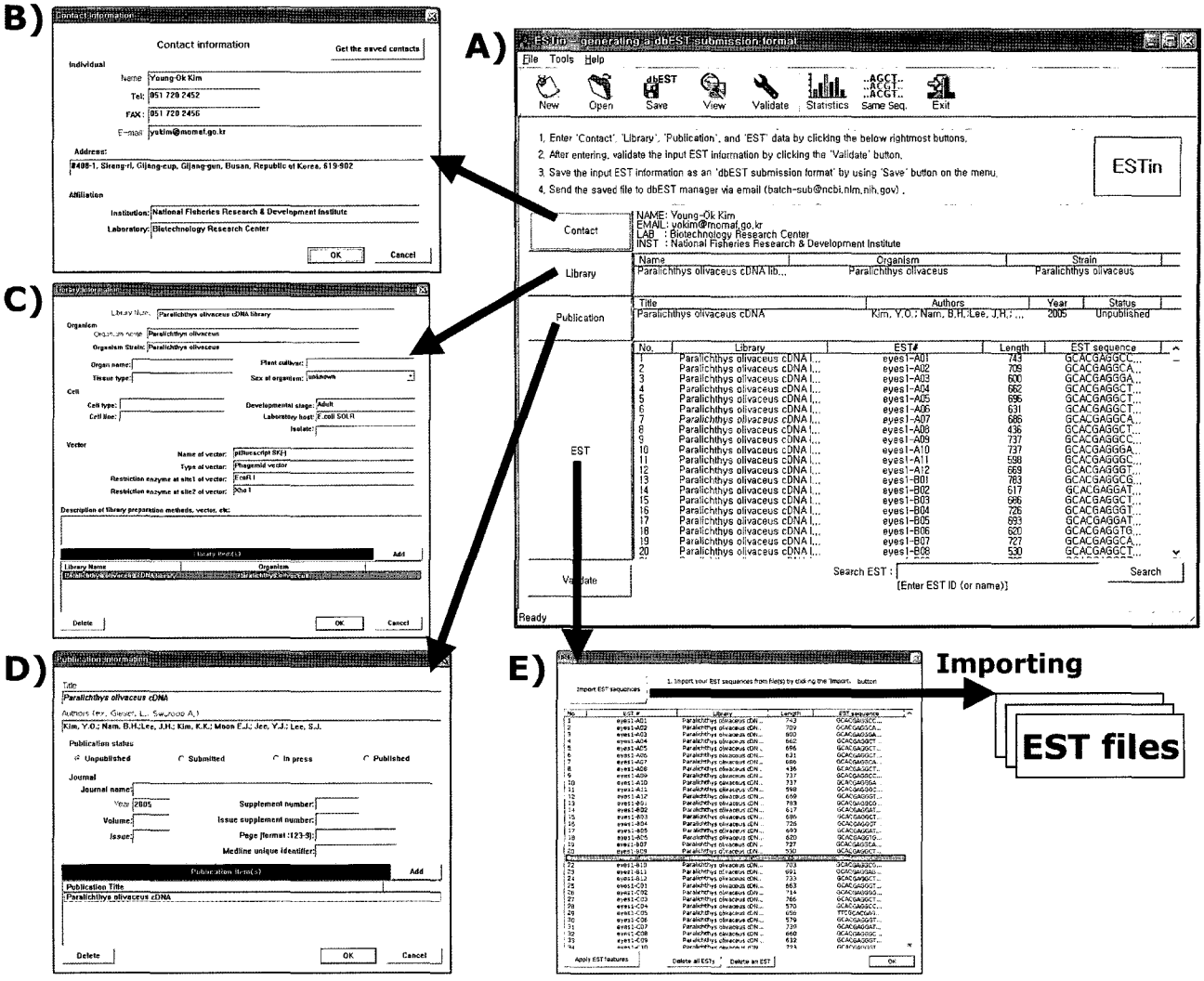

Fig. 1. User interfaces of ESTin. ESTin interface consists of a main window and several input dialogs. Users can input Contact, Library, Publication, and EST data by clicking on the corresponding buttons in the main window. A) Main window. B) Contact input dialog. C) Library input dialog. D) Publication input dialog. E) EST input dialog, in which users import their sequences from EST files.

corresponding EST description line with a bar-separated format. For example, if an EST has four unique features: 'EST0001' for name, 'HHC189' for clone, 'M13 forward' for primer, and '640' for insert length, its description line will be the following: >EST001 |CLONE=HHC189|SEQ_ PRIMER= M13 forward|INSERT=640|. These descriptions are automatically parsed and assigned into each EST fields when imported. ESTin also has a function to allow users to edit each EST feature as well as its sequences. After importing, ESTin checks whether the EST ID is unique and the sequence bases are legal.

ESTin shows useful information for EST sequences such as EST statistics and the same EST sequences. The statistics present mean, standard deviation, and length distribution of all the EST sequences. From the statistics information, the user can detect whether non-EST sequences were included, for example, host genome sequences and vector sequences that are generally much longer than EST sequences. The EST name information that havethe same sequences will help users identify multiple importing of the same ESTs with different names.

ESTin has a validation function to check whether the inputdata is correct against the dbEST format. The main checking points are field format, obligatory fields, and EST references to contact, library, and publication. If an error is detected, the user cannot build a submission file.

Some ESTin functions facilitate building a submission

file. Users can preview the input data as the dbEST format and search an EST during the editing process. ESTin can read a dbEST submission file, which may be used for validating a submission file created by other programs before submitting it to dbEST.

Results and Discussion

ESTin is an easy-to-use tool that allows EST researchers to build submission files with EST sequences and their related data. ESTin's user-friend editing and validating functions will be useful to EST submitters and eventually lead to raising the accuracy of dbEST. The ESTin source code is available to academic users upon request.

Acknowledgements

I am grateful to anonymous ESTin reviewers for their comments. The author thanks Maryana Bhak for editing the manuscript. This work was supported by the Korean Ministry of Science and Technology (under grant number M10437010002-06N3701-00210).

References

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F. *et al.* (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252, 1651-1656.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2006). GenBank. *Nucleic Acids Res.* 34, D16-D20.

Boguski, M.S. (1995). The turning point in genome research. *Trends Biochem. Science* 20, 295-296.

Boguski, M.S., Lowe, T.M., and Tolstoshev, C.M. (1993). dbEST--Database for "Expressed Sequence tags", *Nat Genet.* 4, 332-333.

Mao, C., Cushman, J.C., May, G.D., and Weller, J.W. (2003). ESTAP--an automated system for the analysis of EST data. *Bioinformatics* 19, 1720-1722.

Paquola, A.C., Nishyiama, M.Y., Jr., Reis, E.M., da Silva, A.M., and Verjovski-Almeida, S. (2003). ESTWeb: bioinformatics services for EST sequencing projects. *Bioinformatics* 19, 1587-1588.

Parkinson., J., Anthony, A., Wasmuth, J., Schmid, R., Hedley, A., and Blaxter, M. (2004). PartiGene--constructing partial genomes. *Bioinformatics* 20, 1398-1404.