TBIS: A Web-Based Expert System for Identification of Tephritid Fruit Flies in China Based on DNA Barcode

Zhimei Li [1], Zhihong Li [1,*], Fuxiang Wang [2], Wei Lin [3], Jiajiao Wu [4]

[1]Department of Entomology, China Agricultural University, Beijing, P. R. China

[2] National Agricultural Technology Extension Service Center, Beijing, P. R. China

[3]General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China, Beijing, P. R. China

[4]Guangdong Entry-Exit Inspection and Quarantine Bureau, Guangzhou, P. R. China

[*] Corresponding author, Address: Department of Entomology, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, P. R. China, Tel: +86-10-62731299, Fax: +86-10-62733404, Email: lizh@cau.edu.cn

Abstract. Tephritid fruit flies (Diptera: Tephritidae) include serious agricultural insect pests in the world. Besides causing severe damage to fruits and vegetables, this kind of pests could enter countries or regions with international trade easily. Strict trade quarantine measures are imposed in many countries or regions in order to prevent their introduction and spread. Thus accurate and rapid identification is regarded as an essential component of plant quarantine. Traditional expert systems for assistant identification of agricultural insect pests are based on their morphological characteristics. Compared with the morphological identification, however, molecular identification has more advantages especially for the identification of the immature samples which are intercepted more frequently. Among the molecular identification methods, DNA barcoding is very effective and has been selected by the taxonomists in recent 5 years. In view of the above, a network expert system based on the DNA barcode, Tephritid Barcode Identification System (TBIS) was developed with ASP.NET and C# to improve the molecular identification of fruit fly pests in China. The system was supported by Microsoft SQL server 2008 database. Three functions were provided such as molecular identification based on DNA barcode, information browse and inquiry. DNA sequence similarity alignment dynamic programming algorithm served as the inference mechanism. Molecular identification knowledge was obtained from the public database on the Internet and Plant Quarantine Laboratory of China Agricultural University, which contained about 400 COI sequences of nearly 150 species of fruit flies. Moreover, detailed information such as morphological description and pictures of adult, hosts, and geographical distribution are presented in this system. Mixed with molecular, morphological and distributional data, the system can be used as an identification tool both for quarantine technicians and for educational purposes in China.

Key words: Expert system, Pests identification, DNA barcode, SQL server, Tephritidae

1 Introduction

Fruit flies (Diptera: Tephritidae) include some of the world's most serious agricultural pests. Of the nearly 4,500 species of 500 genera currently known worldwide, over 150 species are considered as pests [1] which cause the loss of billions of dollars in production to a wide variety of fruits, vegetable and flower crops such as citrus, apple, mango, and sunflower. They can easily enter many countries through international trade. At present, over 47 species of fruit flies have been introduced to many countries for purpose or unintentionally by human activities, which lead to a broader distribution of fruit flies. As a result, they limit the development of agriculture in many countries or regions because of the strict trade quarantines imposed to prevent their introduction and spread [2-3]. For example, there are 10 species or genera of fruit flies in the 'list of plant quarantine pests for entry in the People's Republic of China' promulgated in 2007, such as *Anastrepha* Schiner, *Bactrocera* Macquart, *Carpomya incompleta* (Becker), *Carpomya vesuviana* Costa, *Ceratitis* Macleay, *Dacus spp.* (non-Chinese distributed species), *Monacrostichus citricola* Bezzi, *Myiopardalis pardalina* (Bigot), *Rhagoletis spp.* (non-Chinese distributed species) and *Toxotrypana curvicauda* Gerstaecker.

Accurate and rapid identification is important for any pest species and especially necessary for fruit flies because this group of pests like fruit flies occupies an important place in the plant quarantine worldwide. Identification of flies (Diptera), particularly fruit flies, is primarily based on morphological characters of adults. Larvae or pupae intercepted generally need to be raised to adults for identification [1]which necessitate an integrated means to identification of fruit flies that is accurate and rapid for virtual emergency of detecting fruit flies for pest control and management. With advancement in molecular biology and genomic technology molecular approaches offer a effective alternate means for tephritid identification in plant quarantines with greater advantages, such as the superiority to distinguish among related species, complex species, subspecies, biological types, and geographical populations that look alike; can identify all stages of life; can identify a species from bits and pieces; can avoid being affected by individual developmental instars and environmental conditions [4-5]. Therefore researchers are trying to use the molecular biological methods to complete the rapid identification for various instars of fruit flies. Among the molecular identification methods, DNA barcoding is very effective and has been selected by the taxonomists in recent 5 years. DNA barcoding employs sequence diversity in short, standardized gene regions to aid species identification and discovery in large assemblages of life. A 648-bp region of the cytochrome c oxidase I (COI) gene forms the primary barcode sequence for members of the animal kingdom [6-7]. A much smaller fragment MINI COI and ND6 also can be used for species identification. By the way, the Consortium for the Barcode of Life (CBOL) was launched in May 2004 and now includes more than 120 organizations from 45 nations [8].

Expert System is an important branch of Artificial Intelligence, using knowledge and inference to solve problems that only an expert can solve. Originated from the 1960s, through 40 years' development, with rapid expansion of application area, expert systems have been applied to many aspects in plant protection, involving assisted identification of pests, integrated pest management, prediction and forecasting, monitoring and early warning and so on [9-11]. Technology and application of the traditional expert system for pests identification based on morphology have been well developed. With the development of the molecular identification especially the DNA barcoding, it is not only possible but also a trend of biological identification to use the molecular identification due to the advantages of the molecular identification. The Barcode of Life Data System (BOLD) based on

the DNA barcode emerged in time. BOLD is an informatics workbench aiding the acquisition, storage, analysis and publication of DNA barcode records. There are three functional units now available on BOLD such as the Management and Analysis System, the Identification System and the External Connectivity System [8]. Up to now (June 20, 2010), there are 73,592 formally described species with barcodes in BOLD, including Animals, Fungi, Plants and Protists. The number of fruit fly specimens with barcodes is 2,157, distributed in North America, South America, Oceania, Africa, Europe and Asia, among which only 3 specimens in China. There are 517 species of fruit flies with DNA barcodes in the BOLD, however, only 113 species with 344 public sequences which can be downloaded. Considering fruit fly identification in China, the BOLD has limitations as follows: the number of fruit flies without public sequences is 404, almost accounts for 80% of the fruit flies with barcodes; there are no ink drawings which is important for the identification; also there is no basic information such as Chinese name, synonym, host, geographical distribution and morphological characteristics [12]. Up to now, no fruit flies identification systems based on DNA barcode was reported in China.

Considering the reasons mentioned above, Tephritid Barcode Identification System (TBIS) was developed with ASP.NET and C# to improve the identification of fruit fly pests in China. Supported by Microsoft SQL server 2008 database, three functions were provided, such as identification based on DNA barcode, information browse and inquiry. DNA sequence similarity alignment dynamic programming algorithm served as the inference mechanism. Identification knowledge was obtained from the public database on the Internet and Plant Quarantine Laboratory of China Agricultural University, and then represented in the knowledge base of the expert system which contained about 400 COI sequences of nearly 150 species of fruit flies. Moreover, detailed information such as morphological description and pictures of adults, hosts, and geographical distribution were presented in the system.

2 The Knowledge Acquisition

The translation of the knowledge possessed by the experts into a knowledge base is the bottleneck in the process of knowledge acquisition [13-14]. For the system based on molecular biology, knowledge acquisition is also an important problem. By consulting the relevant foreign systems, books and scientific publications, we got the barcode information and basic information of the fruit flies involved in this system.

2.1 Barcode Information Acquisition

As identification knowledge the barcode sequences had been obtained from BOLD and Plant Quarantine Laboratory of China Agricultural University, standardized to certain length, and then represented in the knowledge base of the expert system which contained about 400 COI sequences of nearly 150 species of fruit flies.

2.2 Basic Information Acquisition

Detailed information used to confirm the identification results such as morphological description and pictures of adults, hosts, and geographical distribution were collected from the ‘Identification Atlas for

Important Pest Fruit Flies' wrote by Dr. Jiajiao Wu and the DELTA database (http://delta-intkey.com/ffa/index.htm).

3 TBIS Expert System Design and Development

Three basic functions such as identification based on DNA barcode, information browse and inquiry, together with two secondary functions such as notice and basic knowledge for identification are provided. The general structure which designed according to the functional requirement is shown in Fig.1.

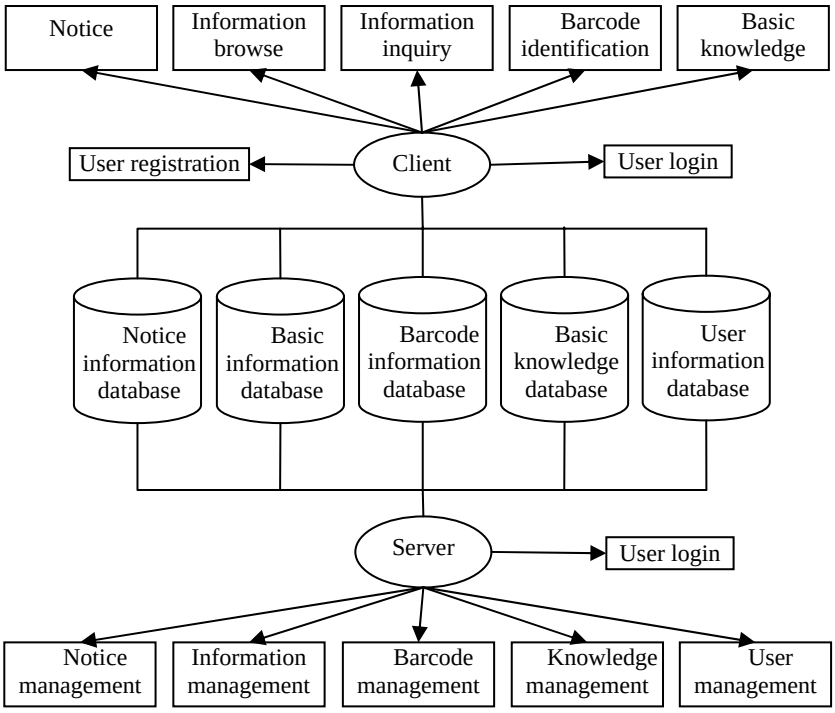

Fig. 1. System structure of TBIS

3.1 Development Software Selection

Since a web-based system can facilitate the data delivery, TBIS was designed to run on the internet to realize the remote access and real-time data sharing. Supported by Microsoft SQL server 2008 database, the system was developed with ASP.NET, C# and HTML. Besides, Dreamweaver 8, Fireworks 8 and Flash 8 were used to design the interface.

3.2 Inference Process

DNA sequence similarity alignment dynamic programming algorithm served as the inference mechanism [15]. Fig.2 shows the inference process. Start from the barcode identification, first, the users should choose the type of the barcodes. At present, there are three types such as COI, mini COI and ND6, and the most common barcode type used in the identification of fruit flies is COI. Second, input the barcode, simply by copy and paste. Third, submit the barcode, the system will show users the

identification result after the similarity alignment between the barcode users input and the barcodes stored in the database by using the DNA sequence similarity alignment dynamic programming algorithm. The identification result is represented in the form of similarity values ordered from high to low. In almost all cases, users can find only one species' similarity beyond 98.00%, and when the similarity value between two DNA sequences is more than or the same as 98.00% we can infer that the two DNA sequences belong to the same species, so it is the final result. However, it is possible that users find two or more species' similarity values beyond 98.00%. For example, when it comes to complex species, you have to do the further identification by the DNAMAN, a kind of software used to create tree map. The barcode sequences of the complex species have been collected into one file folder, users can create a tree map based on the barcode sequences of all the complex species or just choose the similarity values beyond 98.00% ones to save time. Moreover, when there is no species' similarity value beyond 98.00%, maybe it is because the system does not store the species users want to indentify or users' barcode sequence is not a standard one. Here we want to notice users that one species often has more than one barcode sequences from different geographical populations, and these sequences may have different similarity values. At last, users can browse the basic information such as the morphological description and pictures of adults, hosts, and geographical distribution.

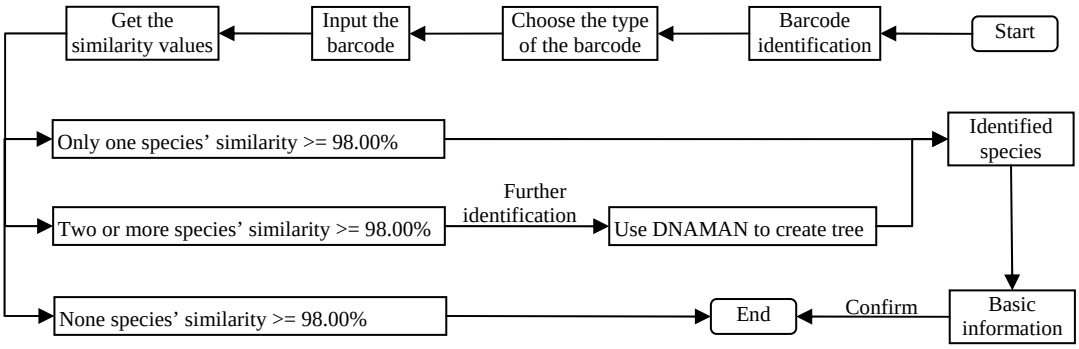

Fig. 2. TBIS inference process

3.3 Users Interface

With abundant combinations of textual information and images, the interface of this system was designed to be easy to operate and user-friendly. There are three main options: information browse, information inquiry and identification based on DNA barcode. On the 'information browse' page, the system displays a list of fruit flies and the user can inspect detailed information related to a specific species after choosing its 'browse' link. If the user clicks the 'information inquiry' option, the system provides an array of inquiry items for users to input keywords. Boolean logical operators ('or', 'and') and inquiry formats ('not exact match', 'exact match') are provided for users to get more suitable results.

As it is shown in Fig.3, the 'identification based on DNA barcode' page firstly offers three barcode types such as COI, MINI COI and ND6 for selecting. If one of the types is selected, the system will access to the database which stores that type of barcodes. Here we chose 'COI' and inputted the COI sequence of the *Bactrocera (Bactrocera) correcta* (Bezzi) collected from Yunnan, China. Fig.4 shows the identification result. Users can further check the detailed information of the identified species by clicking on the Chinese names.

Fig. 3. The identification based on DNA barcode page of TBIS

序号	属名	中文名	学名	相似度数值
结果	果实蝇属	番石榴实蝇	*Bactrocera (Bactrocera) correcta* (Bezzi)	100.00%
2	果实蝇属	番石榴实蝇	*Bactrocera (Bactrocera) correcta* (Bezzi)	99.67%
3	果实蝇属	番石榴实蝇	*Bactrocera (Bactrocera) correcta* (Bezzi)	98.21%
4	果实蝇属	桔小实蝇	*Bactrocera (Bactrocera) dorsalis* (Hendel)	91.22%
5	果实蝇属	菲律宾实蝇	*Bactrocera (Bactrocera) philippinensis* Drew & Hancock	91.22%
6	果实蝇属	桔小实蝇	*Bactrocera (Bactrocera) dorsalis* (Hendel)	91.06%
7	果实蝇属	菲律宾实蝇	*Bactrocera (Bactrocera) philippinensis* Drew & Hancock	91.06%
8	果实蝇属	桔小实蝇	*Bactrocera (Bactrocera) dorsalis* (Hendel)	90.89%
9	果实蝇属	三带实蝇	*Bactrocera (Bactrocera) umbrosa* (Fabricius)	86.67%
10	果实蝇属	三带实蝇	*Bactrocera (Bactrocera) umbrosa* (Fabricius)	86.50%
11	果实蝇属	三带实蝇	*Bactrocera (Bactrocera) umbrosa* (Fabricius)	86.18%
12	果实蝇属	瓜实蝇	*Bactrocera (Zeugodacus) cucurbitae* (Coquillett)	85.04%
13	果实蝇属	瓜实蝇	*Bactrocera (Zeugodacus) cucurbitae* (Coquillett)	84.23%
14	果实蝇属	瓜实蝇	*Bactrocera (Zeugodacus) cucurbitae* (Coquillett)	84.07%
15	果实蝇属	瓜实蝇	*Bactrocera (Zeugodacus) cucurbitae* (Coquillett)	83.41%

Fig. 4. The identification result page of TBIS

4 Conclusions and Future Development

The system was originally conceived to improve the identification of fruit fly pests, the identification of which is very difficult for there are so many species and many of which are extremely similar. Mixed with molecular, morphological and distributional data, the system can be used as an identification tool both for quarantine technicians and for educational purposes in China.

TBIS is different from traditional expert systems based on morphological identification. DNA sequence similarity alignment dynamic programming algorithm served as the inference mechanism. The system contained about 400 COI sequences of nearly 150 species of fruit flies with detailed information such as morphological description and pictures of adults, hosts, and geographical distribution to confirm the identification results in some ways. Moreover, TBIS is a web-based system so it is accessible to Chinese with a computer and internet connection.

Due to the advantages of the molecular identification, it is the trend to use molecular identification with the assistant of the morphological knowledge. However, we should consider the further improvements to make the system more useful. The accuracy of the molecular identification depends on the number of the species and sequences stored in the database as well as how typical and standard the sequences are. Therefore, more species and sequences should be added. What's more, to use different types of DNA sequences at the same time also helps to confirm the identification result, so considering the system' extension, since we have already designed three types of barcodes that can be used for the identification of fruit flies, we need to collect more sequences belong to that three types of barcodes.

Acknowledgements

We would like to thank Prof. Ding Yang, Prof. Zhen Su and Master Jiaqi Liu for the advices of the inference mechanism. We also appreciate to Mr. Wenxin Li for his help on code design. Thank Liang Liang and Yingcai Chen for providing the barcode sequences and advices for the requirement analysis. Thanks for all other members of the Plant Quarantine Laboratory of China Agricultural University (CAUPQL). This study was supported by project of Ministry of Agriculture, China (No. 2009-Z41).

References

1. Zhihong Li, Hanchun Yang, Zuorui Shen: General Animal and Plant Quarantine (in Chinese). China Agricultural University Press, Beijing (2004)
2. Carroll, L.E., White, I.M., Freidberg, A., Norrbom, A.L., Dallwitz, M.J., Thompson, F.C.: Pest Fruit Flies of the World (Version: 15th July 2005). http://delta-intkey.com/ffl/www/_wintro.htm
3. Jiajiao Wu, Fan Liang, Guangqin Liang: Identification Atlas for Important Pest Fruit Flies (in Chinese). Guangdong Science and Technology Press, Guangdong (2009)
4. Fan Liang, Jupeng Zhao, Guangqin Liang, Xuenan Hu: On the Role of the Morphological and Molecular Identification of Insects (in Chinese). Plant Quarantine, vol. 21, pp. 243--244 (2007)
5. Furong Wang, Wei Wu, Defu Rong: Application Prospect of Molecular Biological Techniques in Identification of Quarantine Insect (in Chinese). Journal of Anhui Agri.Sci., vol. 36, pp. 3149--3151, 3162 (2008)
6. Hebert, P.D.N., Cywinska, A., Ball, S.L., deWaard, J.R.: Biological Identifications through DNA Barcodes. Proceedings of the RoyalSociety of London. Series B, Biological Sciences, vol. 270, pp. 313--321 (2003)

7. Savolainen, V., Cowan, R.S., Vogler, A.P., Roderick, G.K., Lane, R.: Towards Writing the Encyclopedia of Life: an Introduction to DNA Barcoding. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, vol.360, pp. 1805--1811 (2005)
8. Ratnasingham, S., Hebert, P.D.N.: BOLD: The Barcode of Life Data System (www.barcodinglife.org). Molecular ecology notes, vol. 7, pp. 1--10 (2007)
9. Zhihong Li, Zuorui Shen, Zhiling Wang: Preliminary Study on the Development and Application of the Expert System for Plant Protection (in Chinese). Plant Protection in the 21st Century, pp. 98--104.China Science and Technology Press, Beijing (1998)
10. Limin Zheng: Principle and Application of Artificial Intelligence and Expert System (in Chinese). China Agricultural University Press, Beijing (2004)
11. Yongli Zheng, Jiaan Cheng, Qianghua Zhang: Expert System and Its Application and Development in Plant Protection(in Chinese). CHINA RICE, pp. 31--33 (2004)
12. Barcode of Life Data System, http://www.boldsystems.org/views/login.php
13. Edward-Jones, G.: Knowledge-based Systems for Pest Management: an Application-based Review. Pestic. Sci., vol. 36, pp. 143--153 (1992)
14. Plant, R.E., Stone, N.D.: Knowledge-based Systems in Agriculture. McGraw-Hill, New York (1991)
15. Wei Li: Introduction to Bioinformatics (in Chinese). Zhengzhou University Press, Henan (2004)