

## Mine Barcode of Life: Information Visualization and Fusion for the Environment and Society

Jen-Hao Cheng, Yang Wang, Paul Yu Chen, Tai-Been Chen, Chun-Jui Chen  
Guan-Cheng Li and Henry Horng-Shing Lu<sup>§</sup>

Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan

### ABSTRACT

Fusing and visualizing information from diverse experimental sources is a key step to integrate complex data on the interactions of biological systems and the Barcode of Life (BOL). There are various sources at numerous repositories, which could be difficult and tedious to access information. The analysis process will be accelerated if a common platform with automatic pipelines exists for information retrieval and fusion. For this purpose, we have developed a web application named *Mine Barcode of Life: Information Visualization & Fusion for the Environment and Society*. Currently, *Mine Barcode of Life* integrates query results from DNA barcodes of mitochondrial sequences, taxonomy trees, images and related information retrieved from different resources, including Consortium for the Barcode of Life, BOL Data System, FishBase, NCBI, Environmental News Service and related websites. This web application has four main features: (1) Identify the species when provided with DNA barcode. (2) Search environment related news via Environmental News Service and Google News. (3) Visualize geographical distribution of different fish species on Google Map. (4) Analyze related fish barcodes with phylogenetic tree. As the progress of Barcode of Life advances, *Mine Barcode of Life* aims to provide an automatic, simple, efficient and publically available web application that makes use of available information to serve all concerned individuals and professionals in the international communities.

**Keywords:** Barcode of Life, *Mine Barcode of Life*, information retrieval, information visualization, information fusion

### INTRODUCTION

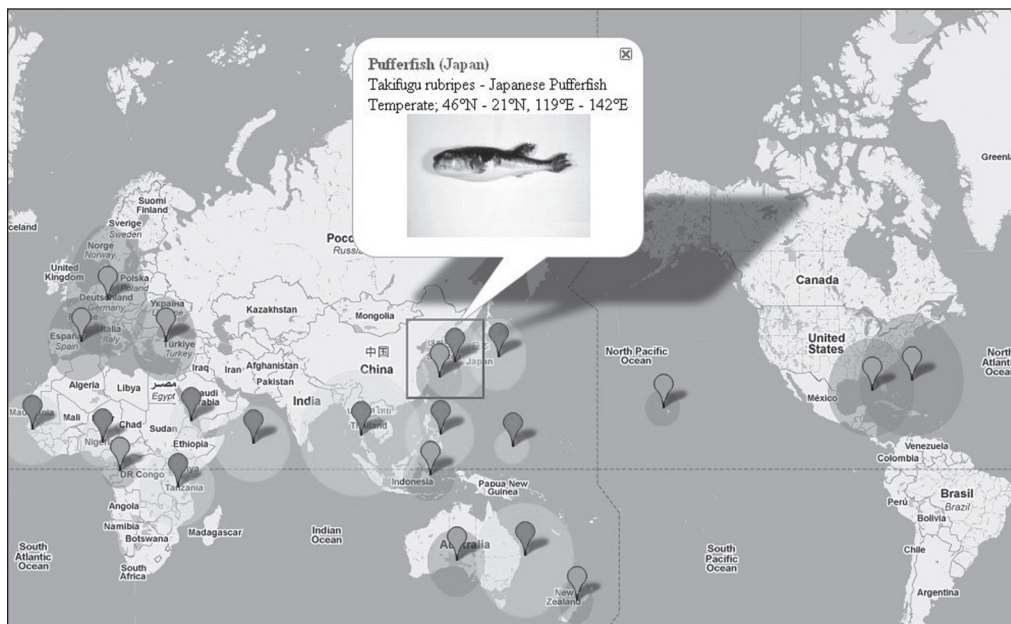
The Consortium for the Barcode of Life (CBOL) is an international initiative aiming to develop DNA barcoding as a standard for the identification of each species [1]. DNA barcoding is a technique that uses a short, standardized DNA sequence as a molecular diagnostic for species-level identification. The DNA barcode sequences are very short, thus they can be obtained quickly and economically. The “Folmer region” at the 5’ end of the Cytochrome C Oxidase subunit 1 mitochondrial region (COI) has emerged as the standard barcode for almost all groups of higher animals [2]. This region is relatively easy to isolate and analyze, for it is around 648 base pairs long and it is flanked by conserved sequences. Studies have shown that COI has low variability within species but a few percent differences across species [2]. Therefore, species

<sup>§</sup>Corresponding author: E-mails: [hslu@stat.nctu.edu.tw](mailto:hslu@stat.nctu.edu.tw)

can be identified with high confidence using COI as DNA barcode.

The motivation for the development of a web application using DNA barcoding was triggered by a FDA warning. In May 2007, the FDA issued a warning not to consume monkfish imported from China, which actually may be poisonous puffer fish [3]. Puffer fish contains tetrodotoxin, a neurotoxin that inflicts paralysis, other serious illness or death. Tetrodotoxin cannot be destroyed by common food processing methods such as cooking or freezing, and there is no antidote [4]. Two people became seriously ill after consuming the puffer fish, which was mis-labeled as the harmless monkfish. These two fish species' distribution overlapped in East China Sea (Figure 1), thus increases the chances of mis-labelling. This unfortunate incident could have been detected and avoided if the DNA barcodes are used and the efficient web application for identifying DNA barcodes exists.

For this purpose, we have developed *Mine Barcode of Life: Information Visualization & Fusion for the Environment and Society* (<http://140.113.114.234/BOL/>). *Mine Barcode of Life* (MBOL) retrieves data from Consortium for the Barcode of Life (CBOL; <http://barcoding.si.edu>), BOL Data System (BOLD; <http://www.barcodinglife.org/>), FishBase (<http://www.fishbase.org/>), NCBI (<http://www.ncbi.nlm.nih.gov/>), Environmental News Service (<http://www.ens-newswire.com/>), and related websites [5-7]. The integrated biological system data, including species identification, taxonomy, images, news and related information, are all inter-connected in MBOL. Moreover, MBOL provides visualization tools for users to perform further analysis at a common platform. Therefore, it simplifies the complications of visiting different sources, downloading the data, and formatting the data for each analysis tool. *Mine Barcode of Life* aims to provide professionals,



**Figure 1:** Geographical Distribution of Puffer Fish and Monkfish

The red marker shows puffer fish; the blue marker shows monkfish. The dialogue box shows a brief description of the puffer fish distributed in East China Sea. The red box identifies the overlapping distribution between puffer fish and monkfish.

government agencies and consumers a simple and efficient way for identifying and analyzing different species, including the illustrative examples in fish species.

## RESULTS

The main purpose of the MBOL is to integrate and visualize information on the biological system. Therefore, the application does not host databases for most of the information retrieved. The general method is to query a specific database, load and present the desired results. MBOL is implemented in JSP. It contains four major components: DNA Barcode Identification, Environment News, Geographical Distribution Visualization, and Phylogenetic Tree Viewer. DNA Barcode Identification matches COI input sequence and returns with the species name; Environmental News includes a news feed of environmental issues and allows news searches; Geographical Distribution Visualization displays an approximate distribution of each fish queries on Google Map; Phylogenetic Tree Viewer allows tree analysis of DNA barcodes. These components merge information from each other and outside databases, and present the information in a user-friendly visualization.

### DNA BARCODE IDENTIFICATION

The DNA Barcode Identification accepts plain text sequence. As the user enters a sequence of DNA barcode, the tool connects to the Reference Barcode Record of BOLD-Identification System (BOLD-IDS) of Barcode of Life Data Systems (BOLD) [8]. BOLD is an online workbench that aids collection, management, analysis, and use of DNA barcodes. BOLD-IDS accepts a DNA sequence and returns a list of species taxonomic assignment. If a taxonomic assignment is found, the MBOL retrieves the top hit species name (e. g. *Takifugu rubripes*). The user can further search this species for news results, NCBI, NCBI Taxonomic Browser, Google, Google News, Google Images, and Geographical Distribution (Figure 2). The inter-connect links allows fusion of information from different experiments and sources, and thus, to further understand the particular biological system.

### ENVIRONMENTAL NEWS

The Environmental News section provides links to news channels. It preloads a RSS news feed for the latest environment news provided by Environmental News Service [9] (Figure 3a). When the user queries this section with *Takifugu rubripes*, MBOL retrieves the related news reports from Google News (Figure 3b). This section is located in the home page, which allows concerned users to receive live news feed as soon as they connect to MBOL.

### GEOGRAPHICAL DISTRIBUTION

Geographical Distribution utilizes Google Map. It allows queries of a common name or scientific name of a species. Upon query, the MBOL retrieves information from FishBase. FishBase is a relational database that contains practically all fish species known to science [10]. MBOL retrieves the latitude and longitude positions from FishBase and mark an approximate geographical distribution of the fish (e.g. *Takifugu rubripes*) on the world map (Figure 4). Clicking on the

● File ○ Text Barcode ID Search:

Go!

(plain sequence only - eg. indel)

**SEARCH RESULT EXAMPLE**

*You searched for barcode id:*

ggagtatcgt	cgaggtatcc	cagctagacc	aaggaagtgt	tgaggggaaga	aggttaggtt
gacaccaatg	aacattactc	cgaagtggat	ttagttcaa	gtgctgtgga	gtgtgtatcc
tgaaaatagt	gggaatcagt	gtacgaatgc	accataaat	gcaatacag	cacccatgga
gaggacgtag	tggaatggg	caactacgta	gtaggtgtcg	tgtaatcga	tgtctagggg
tgaattggct	aggacaattc	cggttaggcc	acccactgta	aataggaaga	tgaagccgag
ggctcatagt	ataggggttt	ctcatttaat	tgatcctcca	tgcaaggttg	caagtcagct
aaatactttg	actcctgtgg	ggatggcaat	aattattgtg	gcagaggtaa	agtaggctcg
ggtgtctacg	tccatgccga	ctgtaaacat	gtgatgggct	catacaataa	aaccaagaag
accgatggcc	attatggctc	agaccatgcc	catgtagccg	aatggttcct	ttttgccgga
gtagtaggct	acgatgtgtg	aaattatccc	gaagccaggg	agaattagaa	tgtagacttc
aggggtgtcca	aagaatcaga	ataagtgttg	gtacaagatg		

**Search output:**  
A result is found...  
Your search returned: **Takifugu rubripes**

- Do a general NCBI search on 'Takifugu rubripes'?
- Do a general cross search on 'Takifugu rubripes' news?
- Taxonomic Browse 'Takifugu rubripes'?
- Google search 'Takifugu rubripes' for common names?
- Google image search 'Takifugu rubripes' for images of specie?
- Show Geographical Distribution of [\[Takifugu rubripes\]](#)?

*Did nothing show up?  
If your search query is too short, the BOLD-IDS system may not pick up any corresponding genus and specie name.*

**Figure 2:** DNA Barcode Identification

DNA Barcode Identification of COI gene of *Takifugu rubripes*

marker will display basic information about the species as well as a link to the FishBase database. Sometimes, a common name is applied to multiple species. Then all of the species will be shown on the map. However, if the database does not contain complete geographical information for the species, a link or links to the FishBase database will be displayed at the top, but geographical distribution will not be displayed (Figure 4). Currently, Geographical Distribution only allows queries on fish species. Furthermore, it does not support multiple species search that displays on the same map.


### PHYLOGENETIC TREE VIEWER

The Phylogenetic Tree Viewer is the only tool that does not link to an outside repository. MBOL employs a locally hosted PhyloWidget application. The PhyloWidget application is a program for viewing, editing, and publishing phylogenetic trees online [11]. Currently, the tree viewer loads the locally-stored COI gene barcodes of two fish taxonomic classes, Actinopterygii and Chondrichthyes (Figure 5). The user can select which class to view the phylogenetic tree. Upon class selection, the tree will be loaded and the user can perform further analysis with the functions of PhyloWidget.

Related News Search:

**Example:** The mislabeling of toxic puffer fish as monkfish.

---

 **Environment News Service**  
 Late-breaking environmental news feed

---

**BASKETBALL STAR YAO MING IS UNEP'S FIRST ENVIRONMENTAL CHAMPION** Posted on :: 08/10/2008 02:56 AM

---

**BEIJING OLYMPICS OPEN BUT AIR DOES NOT CLEAR** Posted on :: 08/10/2008 02:55 AM

---

**TEXAS BIOFUELS WAIVER REQUEST SHOT DOWN** Posted on :: 08/08/2008 01:02 AM

(a) News fed provided by Environment News Service

Related News Search:

**Example:** The mislabeling of toxic puffer fish as monkfish.

---

**Your search results for: *Takifugu rubripes***

**CBSE News: *Takifugu rubripes* (puffer fish) and *Drosophila* ...**  
 You Are Here: Home > News > *Takifugu rubripes* (puffer fish) and *Drosophila melanogaster* (fruit fly) now available on the UCSC Genome Browser ...  
[www.cbse.ucsc.edu/news/article.php?ID=1521](http://www.cbse.ucsc.edu/news/article.php?ID=1521) - 25k - Cached - Similar pages

**Fugu (*Takifugu rubripes*)**  
*Takifugu rubripes* News. Non-coding genes These have been updated for most species, including an miRNA update and HGNC names where possible. ...  
[ensembl.genomics.org.cn/Takifugu\\_rubripes/](http://ensembl.genomics.org.cn/Takifugu_rubripes/) - 14k - Cached - Similar pages

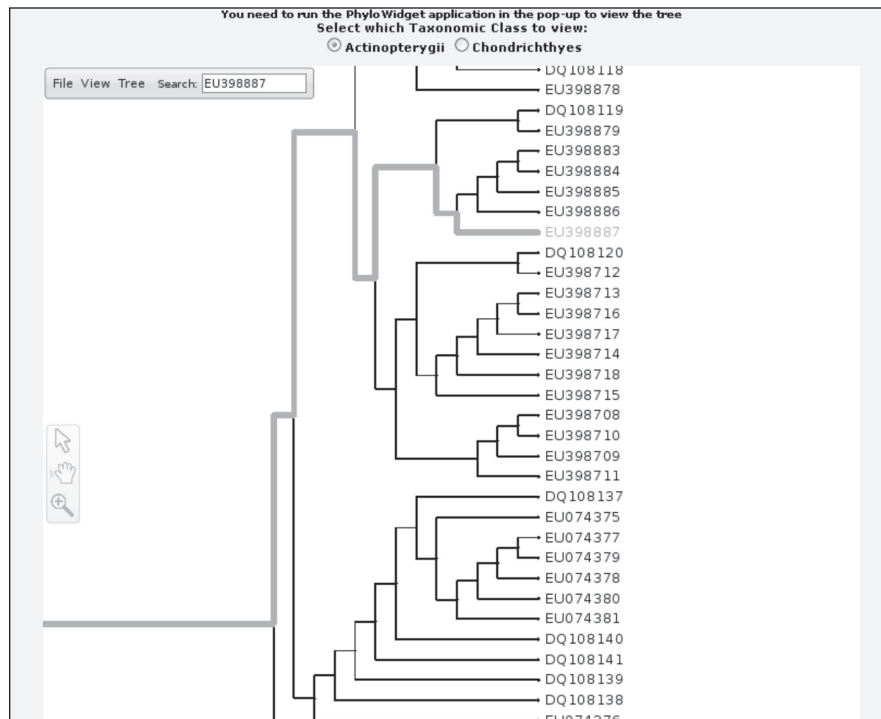
(b) Google News query results on *Takifugu rubripes*

**Figure 3:** Environment News



**Figure 4:** Geographical Distribution

The geographical distribution visualization of Japanese puffer fish



**Figure 5:** Phylogenetic Tree Viewer

The Phylogenetic tree of the COI genes of Actinopterygill

## DISCUSSION

As MBOL links to more data repositories, it aims to provide professionals, government agencies, and consumers a simple and efficient method to collect and visualize information with an automatic pipeline. MBOL provides a perspective in the viewpoint of systems biology. There are vast amount of information gathered from a reductionist approach in understanding biology. However, it is necessary to understand biology in its system using a holistic approach [12]. MBOL provides a potential platform for analyzing the biological system. It allows integrations of reductionist and holistic data, thus presenting all information about an organism in a common platform.

Currently, MBOL is still at its infancy stage with four main features. MBOL was motivated by the mislabeling of puffer fish as monkfish. Therefore, among these four features, Geographical Distribution and Phylogenetic Tree Viewers only provide fish-related information at the current stage. In the future, MBOL plans to offer geographical distribution and phylogenetic data from a variety of databases for different organisms.

Moreover, in order to present the biological system of a particular organism, we plan to incorporate advanced information from more databases, including health and medicine related databases, and high throughput experiment databases. As the field of systems biology progresses, it becomes necessary to study an organism as a whole through integrating data from different sources into a common platform. Combing all the information in a common platform allows a simpler and more efficient process for analyzing the associations among different kinds of data, such as the associations between molecular, cellular, tissue, organism or the interaction with the environment. We imagine a hypothetical scenario that a researcher obtained an unknown tissue sample filled with tetratoxin that could potentially be puffer fish. The research could sequence the COI barcode of the tissue and verify the identity of the tissue through DNA barcode identification, then search for its news and distribution. The researcher could also perform in-depth analysis with phylogenetic trees, health and medicine databases, as well as high throughput experiment databases. Through the analysis, the researcher may be able to develop new relationships by integrating information on the tools incorporated. MBOL aims to provide a platform that transforms this hypothetical scenario to a common practice, where researchers could discover new associations between various information sources.

We also envision values for non-professional usages. The CBOL aims to obtain DNA barcodes for all existing and commercially valuable species. To avoid unfortunate incidents such as the miss-labeling of monkfish as puffer fish, it will be very useful to include the DNA barcode information on consumption food products. MBOL's role can serve as the publically available tool for species identification. Importers, government agents and market places can use MBOL to check the identity of their products in advance before sales and consumptions. Thus, they can ensure the safety of their food products from importation to consumption. Also, this tool is publically available online; consumers can check the validity of their products using wireless internet services during purchases. While the specific details of this future goal are still under development, it has the potential to ensure consumers purchase safe and healthy products from their vendors.

As the progress of Barcode of Life advances, *Mine Barcode of Life* can make use of the information available from various data repositories. MBOL collects information from these repositories into a common platform, and provides visualization and simple analysis for these data with automatic pipelines. Upon entering the DNA barcodes, the user can identify the species names, search for related news, display geographical distributions, and analyze them using phylogenetic tree viewer. The application is simple and efficient, thus it avoids complications regarding collecting information manually from different sources, and reformatting the retrieved data. The tool is also publically available on the internet. Therefore, MBOL serves all concerned users in international communities for both individual and professional needs.

#### ACKNOWLEDGEMENT

We thank the Consortium of Barcode of Life for encouraging this research, providing the sequence identification data, and FishBase for providing detailed information for the each fish species. Finally, we thank National Science Council in Taiwan for providing partial funding for assistants and researches.

#### REFERENCES

- [1] Consortium of Barcode of Life [<http://www.barcoding.si.edu/>]
- [2] Huang, D., Meier, R., Todd, P. A., and Chou, L. M., Slow Mitochondrial COI Sequence Evolution at the Base of the Metazoan Tree and its Implications for DNA Barcoding, *J. Mol. Evol.*, 66(2), 167-74, 2008.
- [3] FDA News [<http://www.fda.gov/bbs/topics/NEWS/2007/NEW01639.html>]
- [4] Hwang, D. F., and Noguchi, T., Tetrodotoxin Poisoning. *Adv Food Nutr. Res.*, 52, 141-236, 2007.
- [5] Google [<http://www.google.com>]
- [6] NCBI Taxonomy [<http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/>]
- [7] BOLD Systems Taxonomy Browser [[http://www.barcodinglife.org/views/taxbrowser\\_root.php](http://www.barcodinglife.org/views/taxbrowser_root.php)]
- [8] Barcode of Life Data Systems [<http://www.barcodinglife.org/>]
- [9] Environmental News Service [<http://www.ens-newswire.com/>]
- [10] FishBase [<http://www.fishbase.org/>]
- [11] PhyloWidget [<http://www.phylowidget.org/>]
- [12] Palsson, B., The challenges of in silico biology, *Nat Biotechnol*, 18(11), 1147-50, 2000.