

BOLD SYSTEMS – narzędzie do identyfikacji gatunków

BOLD Systems to baza danych, w której gromadzone są informacje na temat gatunków występujących na naszej planecie. W myśl idei popularyzacji nauki korzystanie z informacji zawartych w bazie Systemu BOLD jest bezpłatne. Jedyne, czego potrzebujemy, to dostęp do internetu.

Baza Systemu BOLD (z ang. The Barcode of Life Data System) to wirtualna platforma, powstała głównie z myślą o deponowaniu informacji na temat barkodów DNA. Temat genetycznego metkowania organizmów, czyli odczytania dla każdego gatunku unikatowego barkodu umożliwiającego jego identyfikację, stał się na tyle popularny, że doczekał się w 2005 r. własnej bazy danych.

Czym jest?

Baza Systemu BOLD ma do zaoferowania dużo więcej niż surowe dane molekularne. Dzięki niej w łatwy sposób możemy wyszukać informacje taksonomiczne dla 224 tys. gatunków zwierząt, 69 tys. gatunków roślin i 23 tys. gatunków grzybów. Informacje zawarte w bazie BOLD mogą służyć jako doskonały punkt wyjścia w badaniach z zakresu morfologii, systematyki, ekologii czy biologii konserwatorskiej. Znajdziemy w niej zdjęcia okazów zielnikowych, eksponatów muzealnych oraz siedlisk. Dzięki ostatniej, czwartej wersji programu mamy wgląd we wszystkie informacje zdeponowane w bazie – zarówno te opatrzone statusem *public* (informacje udostępnione publicznie), jak również prywatne (*private*), w stosunku do tych ostatnich użytkownicy nie widzą jednak danych dotyczących samych okazów.

210_002 - Flora of Poland [PGLLP]



Przykład rekordu w bazie Systemu BOLD dla próby z gatunku storczyca kulista (*Trausteinera globosa*)

Źródło: www.boldsystems.org (3)

Specimen Details

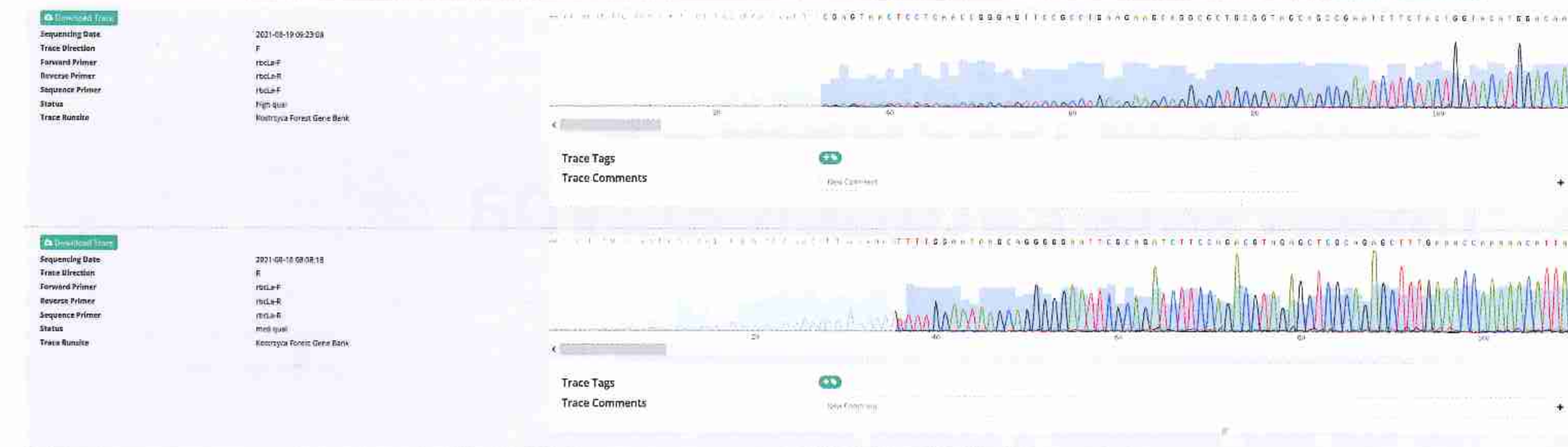
Sample ID:	210_002	Voucher Status:	DNA/Tissue Vouchered Only
Process ID:	PGLLP527-21	Tissue Description:	Fragment of leaf
Project:	PGLLP	Sex:	
Institution Sighting:	Kozłowa Forest Gene Bank	Reproduction:	
Field ID:	210_002	Life Stage:	
Museum ID:		Extra Info:	
Collection Code:		Associated Taxa:	Host: <i>Festuca rubra</i> ; Agrostis capillaris; Hypericum maculatum; Alchemilla sp.; Arnica montana
Reference Link:		Associated Specimens:	
Note:	Voucher Status: no herbarium specimen, seed sample collected		

danych na temat dziesiątków milionów sekwencji wszystkich znanych organizmów na świecie – terminu „barkod”. Od 2017 r., dzięki odpowiedniemu narzędziu BarSTool (z ang. *Barcode Submission Tool*) oznacza się w GenBank sekwencje uznane powszechnie za barkody DNA (markery *COI* dla zwierząt, *ITS* dla grzybów, *rbcL* oraz *matK* dla roślin). BOLD, współpracując z GenBank, pobiera cyklicznie z jego bazy danych informacje na temat wszystkich grup organizmów. Istniejące połączenie między rekordami zapisanymi w bazie BOLD i GenBank wyklucza w stosunku do wielu grup organizmów konieczność korzystania z obu baz danych.

W porównaniu do bazy GenBank, BOLD zapisuje informacje nie tylko na temat sekwencji DNA, ale również surowe dane (z ang. *trace files*) – pliki, które generuje urządzenie do sekwencjonowania w trakcie analizy barkodów DNA. Ma to na celu zwiększenie wiarygodności deponowanych w BOLD sekwencji DNA. Baza archiwizuje również dane terenowe, m.in. datę i miejsce zbioru próby, jej przynależność taksonomiczną, kraj i nazwę repozytorium, a także laboratorium, które odpowiada za wygenerowanie barkodu. Przyjazny dla użytkownika interfejs ułatwia wyszukiwanie informacji w BOLD. Baza daje nam również możliwość eksportowania danych i tworzenia różnego rodzaju zestawień.

Korzystanie z zasobów bazy

Poprzez odpowiednie zakładki, BOLD udostępnia swoim użytkownikom wiele narzędzi, m.in. służących identyfikacji gatunków (zakładka *Identification*) czy określeniu przynależności taksonomicznej (zakładka *Taxonomy*). Poprzez zakładkę *Databases* wyszukamy interesujące nas publikacje naukowe (dostępnych jest ponad 6 tys. tytu-



Dane molekularne

łów), dowiemy się również, jakie jest miejsce pochodzenia (*Geography*) i przechowywania (*Depository*) danej próby. Dzięki specjalistycznym programom (dostępnym w zakładce *Workbench*) baza ma możliwość prezentacji drzewa taksonomicznego wybranego organizmu, określenia dystansu genetycznego w obrębie poszczególnych jednostek systematycznych czy eksportowania interesujących nas informacji w formacie arkusza kalkulacyjnego. Każdy rekord w BOLD, przedstawiający informacje na temat danego okazu, składa się z nagłówka z numerem próby (z ang. *field ID*) i nazwą projektu, zdjęć okazu, szeregu danych dotyczących identyfikacji okazu, miejsca jego przechowywania, typu tkanki, fazy wzrostu, płci, współwystępujących taksonów (pole *Specimen Details*). Rekord przedstawia również aktualne ujęcie taksonomiczne oraz szczegółowe dane terenowe okazu – zarówno w wersji tekstowej, jak i graficznej. W odniesieniu do danych sekwencyjnych, każdy rekord archiwizuje informacje na temat danej próby, dla której zdeponowano w bazie barkody DNA, podaje ich nazwy, długości, sekwencje nukleotydowe i aminokwasowe (zarówno w wersji tekstowej, jak i graficznej). Oceniana jest również jakość sekwencji barkodów. BOLD umożliwia także dodawanie komentarzy (hasła kluczowych), które mogą być wykorzystywane do tworzenia dodatkowych opisów.

Identyfikacja gatunków

Istnieją dwa sposoby szukania informacji w bazie systemu BOLD. Jeden polega na wyszukiwaniu tekstowym za pomocą słów kluczowych, np. nazwy gatunku, nr. próby, kraju pochodzenia, nazwy zbieracza, depozytorium itp. Drugi wykorzystuje sekwencje barkodów DNA, szukając podobnych sekwencji w bazie. W przypadku identyfikacji gatunków za pomocą barkodów DNA, plik wynikowy

przedstawia listę 100 gatunków (pierwszy z nich oznaczony jest jako *Top Hit*), dla których algorytm BOLD znalazł najbardziej podobne do szukanej sekwencji barkodów (*COI* dla zwierząt, *ITS* dla grzybów, *rbcL* bądź *matK* dla roślin). Wyszukane sekwencje analizuje się pod kątem wskaźnika BEST ID. Im wyższa wartość, tym większe podobieństwo między badanymi sekwencjami.

Wymagania dotyczące deponowania informacji

Jednym z głównych problemów bioinformatycznych istniejących baz danych jest możliwość deponowania w nich sekwencji niskiej jakości. W przypadku bazy BOLD opis każdego rekordu podlega szczegółowej weryfikacji przez specjalistów z Uniwersytetu Guelph. Spójny i jednoznaczny system nazewnictwa gatunków, a także konieczność zdeponowania surowych danych sekwencyjnych, zapewnia wysoką jakość zdeponowanych w bazie informacji.

Każdy użytkownik bazy BOLD po zalogowaniu dostaje możliwość założenia własnego projektu i deponowania w nim informacji – terenowych, taksonomicznych, molekularnych oraz cyfrowych (w postaci zdjęć okazów) poprzez moduł WORKBENCH. Jednorazowe zamieszczanie dużej ilości informacji odbywa się za pomocą protokołów (.xls) i skompresowanych plików cyfrowych. W chwili, gdy rekord zostaje wprowadzony do BOLD, system nadaje mu unikatowy numer, tzw. *Process ID*. Numer ten niesie również informacje na temat akronimu projektu oraz roku zamieszczenia informacji w bazie Systemu BOLD.

Rekord dotyczący barkodu DNA nie uzyska statusu formalnego barkodu dopóki nie spełni minimalnych wymagań, tj. podania m.in. nazwy gatunku, instytucji przechowującej próbę, danych dot. zbioru,

osoby identyfikującej próbę. W przypadku barkodów zwierzęcych i roślinnych nie są przyjmowane sekwencje o długości poniżej 500 nukleotydów.

Polski wkład

W bazie BOLD znajdziemy informacje na temat 1 176 gatunków zebranych na terenie Polski. Wśród 6 725 okazów, 86% przypisanych jest do gatunku (5 791 prób). Wśród nich znajdują się gatunki z rodzaju: *Homo* (845 rekordów), *Aceria* (680), *Gammarrus* (275), *Argiope* (168), *Aneura* (166), *Dikerogammarus* (117), *Chaetopteryx* (92), *Rhithropanopeus* (84), *Abacarus* (64) oraz inne (3369) i nieokreślone (865). Jednostkami umieszczającymi informacje na temat prób z Polski są nie tylko krajowe instytucje. Informacje te pochodzą również z Uniwersytetu Guelph w Kanadzie, Uniwersytetu w Minnesocie w Stanach Zjednoczonych (University of Minnesota Insect Collection) czy Zoologische Staatssammlung München (Niemcy). 4880 rekordów zaimportowano z bazy danych GenBank. Wśród jednostek krajowych, największy wkład w deponowanie informacji na temat gatunków zwierząt posiada Uniwersytet Łódzki, Katedra Zoologii Bezkręgowców i Hydrobiologii (275 zdeponowanych okazów). W odniesieniu do gatunków roślin, największą kolekcję zdeponował Leśny Bank Genów Kostrzyca (104 gatunki, 388 okazów, 718 sekwencji barkodów DNA).

W związku z prowadzeniem na dużą skalę międzynarodowych projektów barkodowania DNA, skatalogowanie wszystkich organizmów na świecie wydaje się kwestią czasu. Nie podlega więc dyskusji, że zasoby Systemu BOLD będą się stale powiększać. Dotyczy to również liczby użytkowników.

Ewa Kaczmarek

Więcej szczegółów na temat bazy danych BOLD SYSTEMS na stronie www.boldsystems.org.