Journal of Biotechnology 261 (2017) 187–193

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Mobilization and integration of bacterial phenotypic data—Enabling next generation biodiversity analysis through the Bac*Dive* metadatabase

Lorenz C. Reimer, Carola Söhngen, Anna Vetcininova, Jörg Overmann*

Leibniz Institute DSMZ – German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany

ARTICLE INFO

Keywords:
Metadata
Bacterial biodiversity
Bacterial phenotype
Data mobilization
Database

ABSTRACT

Microbial data and metadata are scattered throughout the scientific literature, databases and unpublished lab notes and thereby often are difficult to access. Hot spots of (meta)data are internal descriptions of culture collections and initial descriptions of novel taxa in primary literature. Here we describe three exemplary mobilization projects which yielded metadata published through the prokaryotic metadatabase Bac*Dive*. The Reichenbach collection of myxobacteria includes information on 12,535 typewritten index cards which were digitized. A total of 37,156 data points were extracted by text mining. In the second mobilization project, Analytical Profile Index (API) tests on paper forms were targeted. Overall 6820 API tests were digitized, which provide physiological data of 4524 microbial strains. Thirdly, the extraction of metadata from 523 new species descriptions of the *International Journal of Systematic and Evolutionary Microbiology*, yielding 35,651 data points, is described. All data sets were integrated and published in Bac*Dive*. Thereby these metadata not only became accessible and searchable but were also linked to strain taxonomy, isolation source, cultivation condition, and molecular biology data.

1. Introduction

The analysis of microbial diversity provides the basis for the understanding of ecosystem functions and of microbial evolution. At the same time, microbial diversity offers innovative solutions for biotechnology, agriculture, and public health. Over the past two decades, advances in molecular biology have allowed the collection of a large number of microbial nucleotide and protein sequences which now populate the public databases (SILVA, Quast et al., 2013; GenBank, Cole et al., 2014; RDP, Agarwala et al., 2016). So far, only a limited amount of contextual data of molecular sequences is available. The MIMARKS and MIxS standards were established six years ago (Yilmaz et al., 2011) and include mostly information on the type and geographic location of the environment a particular sequence originated from. In parallel to the rapid growth of the sequence databases, over 600 novel species of Bacteria and Archaea have been described yearly over the past decade (Overmann, 2013) and this number has even been increasing lately to 800 novel species per year (Overmann et al., 2017). This has created a considerable and increasing amount of taxonomic, morphological, biochemical, and phenotypic information which is often distributed across different sources. Most importantly, these taxon-associated data are not linked to each other or to the existing molecular sequence data. However, taxon-associated data need to be readily available and searchable for different types of studies such as (i) searching for bacteria with a particular phenotype, (ii) comparison of the physiological characteristics of a bacterium with those predicted from *in silico*-analyses of genome sequences, or (iii) predicting the physiology of not-yet-cultured types of bacteria based on data existing for close relatives.

Published taxon-associated data are distributed across different types of research publications, appear in different non-standardized formats, and therefore are usually difficult to access. In addition, many data that are produced every day are never published but remain hidden as laboratory notes all over the world. Particularly large aggregations of these hidden data can be found in public or private bacterial culture collections. Here, data are often collected over years in a systematic fashion and for thousands of microbial strains, without considering publication. The present contribution describes the strategy developed for the mobilization and aggregation of large, unpublished and non-standardized datasets and their integration into the bacterial metadatabase Bac*Dive*, and analyzes the opportunities and challenges of this approach.

2. Material & methods

12,535 index cards from the Reichenbach collection were scanned

* Corresponding author at: Leibniz-Institut DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, Inhoffenstraße 7 B, 38124 Braunschweig, Germany.
E-mail address: joerg.overmann@dsmz.de (J. Overmann).

http://dx.doi.org/10.1016/j.jbiotec.2017.05.004
Received 13 February 2017; Received in revised form 1 May 2017; Accepted 4 May 2017
Available online 06 May 2017

with a resolution of 300 dpi in greyscale. The optical character recognition was carried out using the "Omnipage 18 professional" software package. To maximize the yield the files were processed with three different settings for brightness (bright, middle, dark) and the according optimal contrast adjustment. 2478 photographic slides were scanned with a Nikon Super Coolscan 5000 ED with 48Bit HDR and a resolution of 4000 dpi. The three OCR results were processed by regular expressions (supplemental Table 1) and returned positive, negative or no result for the analyzed terms. The results were joined afterwards as follows: Combinations of positive and no result were rated as positive. Combinations of negative and no result were rated as negative. Combinations involving positive and negative results occurred very rarely and were solved by manual inspection of the original index card.

Key – value pairs determined by text mining of index cards from myxobacteria were inserted by SQL scripts into the Bac*Dive* MySQL database. Data were introduced into the fields "Culture medium name", "Culture medium composition", "Culture medium growth", "Enzyme", "Enzyme activity", "EC number", "Observation", "Metabolite", "Antibiotic sensitive/intermediate/resistant", "Antibiotic sensitive/intermediate/resistant concentration", "Sample type/isolated from", "Geographic location" and "Country/Continent" finally being visible in the sections "Culture and growth conditions", "Morphology and physiology" and "Isolation, sampling and environmental information" on Bac*Dive*'s web pages.

For API test data the database scheme, as well as the web interface, was adapted to allow for the display of API tables within the section "Morphology and physiology". For this purpose, the Bac*Dive* database was extended by 547 data fields to take up the data for 16 different API test types (API 20 A, API 20 E, API 20 NE, API 20 STRP, API 50 CH assimilation, API 50 CH acid, API CAMPY, API CORYNE, ID 32 E, ID 32 STAPH, API LISTERIA, API NH, rapid ID 32 A, rapid ID 32 STREP, API STAPH, API ZYM).

Data from species descriptions in IJSEM were collected within spreadsheets comprising 152 data fields. Data were standardized by applying controlled vocabulary and whenever feasible identifier like EC numbers for enzymes, ChEBI IDs for metabolites was assigned. Data were exported in a database compatible format and imported into the Bac*Dive* database.

3. Results and discussion

3.1. BacDive – the bacterial metadatabase

In 2012 Bac*Dive* has started mobilizing metadata from internal descriptions of the German culture collection DSMZ (Leibniz Institut DSMZ – Deutsche Sammlung für Mikroorganismen und Zellkulturen) and from the collection of Myxobacteria of Prof. Dr. Hans Reichenbach, who gathered data at the Gesellschaft für Biotechnologische Forschung (GBF, which is now the Helmholtz Centre for Infection research) for more than 6000 myxobacterial strains. Both sources were previously not publicly available and served as a starting point to build up the bacterial metadatabase. In the following years, Bac*Dive* continuously extended not only the strain repertoire but also data content by integrating the *Compendium of Actinobacteria* (https://www.dsmz.de/bacterial-diversity/compendium-of-actinobacteria.html), compiled by Dr. Joachim Wink and the descriptions of the Jena Microbial Resource collection (JMRC). These projects have so far resulted in an accumulated number of 373,628 metadata points for 54,610 bacterial and archaeal strains provided by Bac*Dive* (Release 09/2016).

Data are collected within 749 active data fields and assigned to the seven thematic sections "*name and taxonomic classification*", "*morphology and physiology*", "*culture and growth conditions*", "*isolation, sampling and environmental information*", "*application and interaction*", "*molecular biology*" and "*strain availability*". Within the graphical user interface (GUI), data can be accessed either using *the simple search* or the *advanced search*. The *simple search* function enables a strain selection by searching for species names or culture collection identifiers. The *advanced search* function enables the user to perform comprehensive queries by searching and combining 61 data fields. Data can be retrieved within the GUI using the customizable CSV-export function.

The academic and non-commercial use of information retrieved from Bac*Dive* is free under the condition that the source is indicated. However, redistribution and commercial use require permission.

3.2. Mobilization of the Reichenbach collection of myxobacteria

The first mobilization project targeted the Reichenbach collection of Myxobacteria, which was handed over to the Leibniz Institute DSMZ in the year 2000. This collection of 7565 myxobacterial strains was accompanied by 12,535 index cards with descriptions, 2478 microscopic pictures, and 892 handwritten media recipes. The texts on the index cards were written with a typewriter in German language and comprised data about the taxonomy, isolation source, cultivation, physiological properties and morphological description. Non-standard abbreviations had been used for the strain descriptions.

The cards were digitized and an optical character recognition (OCR) was performed using three different settings of brightness and contrast. The high frequency of non-standard abbreviations and specialized terms rendered an automated translation impossible. Therefore a custom strategy was developed to extract the relevant information. While the isolation sources were manually translated, the remaining text was screened for conserved terms as targets for text mining. By this, 80 target terms were determined which contribute information about cultivation media, polymer degradation ability, enzymatic activity and antibiotic resistance information (detailed descriptions can be found in the supplemental Table 1). After establishing regular expressions for these terms, data were extracted from the 12,535 index cards. To maximize the yield and to minimize errors introduced by the OCR this extraction was applied to the three different OCR data outputs. For the single OCR results 32,705, 33,047 and 35,346 data points were gained. By joining these results, a total of 37,156 data points could be generated, and hence an increased yield of data points between 5.1–13.6% be achieved. This demonstrates the benefit of performing the OCR under different conditions. To examine the precision (true positives/(true positives + false positives)) and recall (true positives/(true positives + false negatives)) of the text mining, a test data set of 500 randomly selected index cards was manually reviewed. 378 of these index cards contained targets for the text mining and these targets were determined with a precision of 99.45% and a recall of 96.03%. The high precision can be traced back to a low number of false positives which is ensured by a conservative formulation of the regular expression. The lower recall value is caused by false negatives that can be traced back to the high diversity of terms used by the authors of the index cards and to errors which were introduced by the OCR.

For the 12,535 index cards, a total of 9607 were found containing text mining targets. While the highest number of targets per single index card was 17, the distribution shows that the most prevalent number of targets per card was four (2682), followed by three targets (2389) per card (Fig. 1). The distribution of the occurrence of text mining targets shows that with 7003 occurrences the most frequent term is *Yeast lysis* (the ability of the myxobacteria to lyse yeast cells) followed by 6240 occurrences of the term *vy/2* (denoting growth on the cultivation medium vy/2) (Table 1). The list of targets is dominated by information on cultivation media (M) but also contains information about lysis and polymer degradation (D) as for example decomposition of chitin (1576 entries) and of casein (928 entries). To a lower extent, the targets provide antibiotic resistance information (A), for example, toli/Kan4 (804), Kan5 (694 entries) or Kan17 (490 entries), describing the growth on different media containing the antibiotic kanamycin, sometimes combined with up to three other antibiotic substances. Comparably low numbers of entries were found for the enzymatic activity tests (E) for cytochrome oxidase (85 entries) and catalase (64