Titel des Vorhabens: DiASPora

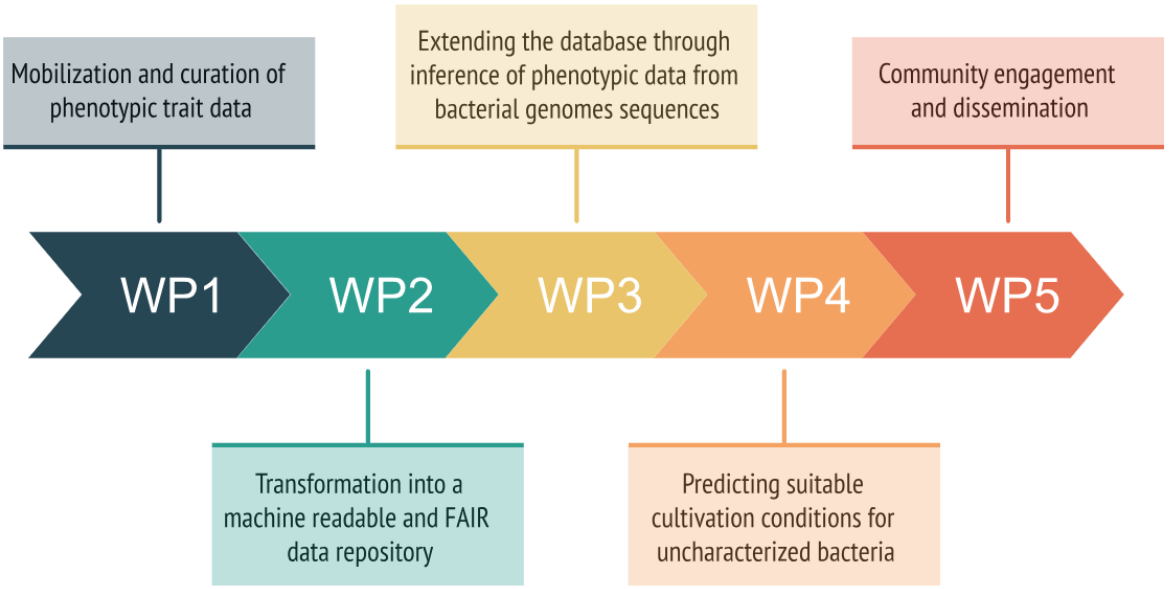
Projektnummer/Aktenzeichen: K280/2019

Executive Summary

Im Rahmen des DiASPora-Projektes konnten wichtige Meilensteine für die Bereitstellung und Vorhersage mikrobiologischer Daten erreicht werden. Die semi-automatische Extraktion solcher Daten aus der mikrobiologischen Primärliteratur konnte durch KI-basierte Text Mining-Ansätze beschleunigt und verbessert werden. Die im Projekt entstandene Annotationsplattform wird an der DSMZ für die Annotation der Literatur eingesetzt und wird in zukünftigen Projekten ausgebaut und erweitert werden. Weiterhin wurden die phänotypischen Daten durch KI-gestützte Methoden weiter angereichert. Random Forest-Modelle wurde auf Basis von Genomannotationen trainiert und auf einen großen Datensatz von fast 20.000 Stämmen aus der Bac*Dive*-Datenbank angewendet. Die entstandenen Modelle wiesen eine überragende Performance auf den Test-Datensätzen auf und hatten eine Genauigkeit von mehr als 95 %. Das Ergebnis sind mehr als 55.000 komplett neue Datenpunkte, die mit Evaluierungsmetriken und Wahrscheinlichkeiten in Bac*Dive* veröffentlicht werden konnten. Diese und weitere Modelle wurden in eine Plattform integriert, die sich der Vorhersage von Wachstumsbedingungen widmet. Um auch Kulturmedien erfolgreich vorhersagen zu können, wurden mehr als 3.200 Medien der DSMZ, aber auch von anderen Kultursammlungen, standardisiert und in eine relationale Datenbank überführt. Das Ergebnis ist die Kulturmediendatenbank Media*Dive*, die der Öffentlichkeit zur Verfügung gestellt wurde und in den vergangenen 12 Monaten bereits von durchschnittlich 5000 Nutzer:innen pro Monat besucht wurde. Auf Basis der standardisierten Kulturmedien konnte die Plattform um eine KI-gestützte Methode erweitert werden, die Medien für zurzeit unkultivierte Mikroorganismen vorhersagen kann. Diese Plattform wird zurzeit in Laborexperimenten getestet. Zu guter Letzt wurde eine umfassende Ontologie für alle im Projekt verwendeten Daten geschaffen. Diese wurde mit bereits existierenden Ontologien abgestimmt. Auf Basis der Ontologie und mithilfe des RDF-Frameworks konnte die Bac*Dive*-Datenbank inklusive der durch Text Mining gewonnenen Literatur und der vorhergesagten Daten in einen Maschinen-lesbaren Wissensgraphen übersetzt werden. Dieser soll in Zukunft noch weiter ausgebaut werden, um neuartige Datenverknüpfungen aufzuzeigen. Hier soll unter anderen eine Anknüpfung an Initiativen der Partnereinrichtungen, wie DSMZ *Digital Diversity* und NFDI4Microbiota, für Mehrwert sorgen.
Trotz anfänglicher Schwierigkeiten bei der Personalfindung und dem eingeschränkten persönlichen Kontakt durch die Corona-Pandemie konnte das Projekt innerhalb des verlängerten Projektzeitraumes vollständig abgeschlossen werden. Es konnten drei Publikationen veröffentlicht werden, zwei davon in anerkannten Fachzeitschriften (peer-review) und eine weitere, die eher die interessierte Fachöffentlichkeit adressiert. Außerdem konnten moderne Software und KI-Modelle im Rahmen des Projektes entwickelt werden, die unter OSI-konformen Open Source Lizenzen veröffentlicht wurden, was eine Nachnutzbarkeit für weitere Projekte garantiert.

1. Zielerreichung und Umsetzung der Meilensteine

Der Arbeitsplan war in fünf Arbeitspakete unterteilt, darüber hinaus wurden keine Meilensteine festgelegt.



Ziel des ersten Arbeitspaketes war die Extraktion phänotypischer Daten aus der mikrobiologischen Literatur durch großangelegtes Textmining unter Anwendung von Techniken der künstlichen Intelligenz (KI), die durch das Feedback von mikrobiologischen Kuratoren trainiert werden. Für die Entwicklung und Etablierung des Textmining-Workflows wurden zehn relevante Zeitschriften und mehr als 6000 in der Bac*Dive*-Datenbank dokumentierte Artikel ausgewählt. Für das automatische Beziehen von maschinenlesbaren Artikel-Versionen wurde ein Workflow entwickelt, der abhängig von der Verfügbarkeit auf PubMedCentral zurückgriff oder die Inhalte durch automatisches Web-Scraping bezog. Wir erstellten verschiedene KI-basierte Annotationsmodelle unterschiedlicher Komplexität zur Extraktion phänotypischer Eigenschaften und Wachstumskonditionen mikrobieller Spezies und wählten das beste Modell aus, das mit 269 Artikeln trainiert wurde. Im Vergleich mit manuellen Annotationen von Expert:innen zeigte das Modell eine hohe Trefferquote von 94%. Das Modell wurde in ein Web-basiertes Annotations-Tool integriert, welches die Annotationszeit eines Volltextartikels von über 30 Minuten auf etwas mehr als 10 Minuten reduziert.

Ziel des zweiten Arbeitspaketes war es, mit Hilfe des *Resource Description Framework* (RDF) die Daten der Bac*Dive*-Datenbank in ein maschinenlesbares und verarbeitbares Format umzuwandeln und diese Daten anschließend zur Erstellung eines Wissensgraphen zu verwenden. Um diese Datenbank zu semantifizieren, wurden folgende Produkte entwickelt:
a) Die Bac*Dive*-Ontologie, mit der eine zentralisierte und umfassende Ressource für Daten zur bakteriellen Vielfalt geschaffen werden soll. Die Schaffung einer umfassenden Ontologie erfüllt nicht nur die unmittelbaren Bedürfnisse von Forschenden, sondern öffnet auch den Weg für breitere Anwendungen in Bereichen wie Biotechnologie, Bioökonomie, menschliche Gesundheit und Umweltschutz;
b) eine Transformationspipeline, die entwickelt und eingesetzt wurde, um die relationale Datenbank als Eingabe zu akzeptieren, die fehlenden bzw. falsch-formatierten Daten so weit wie möglich zu validieren und zu korrigieren, die Bac*Dive*-Ontologie auf die Daten anzuwenden, die RDF-Tripel zu erzeugen und sie als Wissensgraph (engl. Knowledge Graph) zu speichern; und
c) der Knowledge Graph (KG), der auf GraphDB, der Bac*Dive* Ontologie und SPARQL Endpunkten aufbaut. Benutzer:innen können eine bestimmte Version der Bac*Dive*-Datenbank auswählen und diese Version mit der vollen Leistungsfähigkeit von SPARQL abfragen. Im Laufe des Projekts hat die TIB eine Reihe sogenannter Kompetenzfragen (Referenz-Nutzerabfragen) in SPARQL formuliert, um die Aussagekraft der KG zu demonstrieren. Diese Referenzabfragen werden von DSMZ und ZB MED genutzt, um sie in ihre Vorhersagealgorithmen zu integrieren.

Ziel des dritten Arbeitspaketes war es, phänotypische Vorhersagen aus (Meta-)Genomdaten abzuleiten und damit die Bac*Dive*-Datenbank anzureichern. Es konnten mehr als 55.000 Genome zu prokaryotischen Stämmen aus Bac*Dive* assoziiert werden (Reimer *et al.* 2022). Diese wurden mit Pfam annotiert und als Basis genutzt, um mehrere KI-gestützte Modelle zur Vorhersage von Phänotypen zu trainieren. Schließlich wurden die besten Modelle ausgewählt, um die Bac*Dive*-Datenbank mit synthetischem Wissen anzureichern. Es wurden 55.000 völlig neue Datensätze für fast 20.000 Stämme hinzugefügt und 38.500 bestehende Datensätze wurden mit einer Übereinstimmung von mehr als 95 % geprüft. Sowohl die Vorhersagen als auch die zu Grunde liegenden Trainingsdaten wurden über die Bac*Dive*-Datenbank veröffentlicht und ein Manuskript dazu befindet sich gerade in Vorbereitung.

Ziel des vierten Arbeitspaketes war es, Kultivierungsbedingungen und Wachstumsmedien auf Basis der bereits genannten Daten und mittels KI-gestützter Methoden vorherzusagen. Dazu wurden die vorhandenen Kulturmedien der DSMZ aus unstrukturierten Dokumenten (PDF und Word) extrahiert und standardisiert. Die daraus resultierten Medienrezepte wurden manuell kuratiert, in ihre finale Komposition übersetzt und mit weiteren Wachstumsdaten, sowie Kulturmedien aus anderen Kultursammlungen (JCM, CCAP) angereichert. Das Ergebnis ist die umfassende Kulturmediendatenbank Media*Dive*, die der Öffentlichkeit zur Verfügung gestellt (https://mediadive.dsmz.de) und hochrangig publiziert wurde (Koblitz *et al.* 2023). Darüber hinaus wurde ein Prototyp zur Vorhersage von Kulturmedien auf Basis von kNN-Modellen entwickelt. Dieser konnte bereits vielversprechende Ergebnisse vorweisen, die vorgeschlagenen Kulturmedien werden zurzeit aber noch von Kurator:innen der DSMZ im Labor geprüft.

Ziel des fünften Arbeitspaketes war es, die im Projekt verfügbar gewordenen Daten der Öffentlichkeit zur Verfügung zu stellen und mit den Forschenden in Kontakt zu treten. Aufgrund der Corona-Pandemie konnten keine Onsite-Workshops stattfinden und die Nutzerumfragen wurden nur in beschränktem Ausmaß durchgeführt. Ansonsten konnte das Arbeitspaket vollumfänglich umgesetzt werden. Der Quellcode der entstandenen Software wurde der Öffentlichkeit unter OSI-zertifizierten Open Source-Lizenzen zur Verfügung gestellt: die Text Mining-Modelle (WP1) sind in GitHub (https://github.com/foerstner-lab/DiASPora_IllumINAte) veröffentlicht; die Bac*Dive* Ontologie und die Semantifizierungs-Pipeline (WP2) sind in GitHub (https://github.com/TIBHannover/diaspora/tree/V1.1.0/wp2) veröffentlicht; und die Modelle zur Phänotypvorhersage (WP3) sind ebenfalls in GitHub (https://github.com/JKoblitz/bacdive-ai) veröffentlicht. Die resultierenden Ergebnisse sind in Bac*Dive* öffentlich gemacht (CC-BY) und durch eine Referenz zum DiASPora-Projekt gekennzeichnet (https://bacdive.dsmz.de). Die Kulturmedien und Wachstumsdaten (WP4) sind in der neuen Media*Dive*-Datenbank ebenfalls unter der CC-BY 4.0 Lizenz veröffentlicht (https://mediadive.dsmz.de). Community Engagement fand in zwei Online-Workshops zu Bac*Dive* und Media*Dive* statt, die mit 30 bis 50 Teilnehmenden gut besucht waren.

2. Aktivitäten und Hindernisse

Die durchgeführten Aktivitäten und damit verbundene Hindernisse wurden bereits im vorherigen Punkt *Zielerreichung* ausgeführt. Zusätzlich haben sich die Projektinstitutionen alle 3 Monate über den Fortschritt des Projektes ausgetauscht und weitere Arbeiten aufeinander abgestimmt. Regelmäßigere Treffen zu Einzelthemen fanden nach Bedarf statt, wodurch während der gesamten Projektlaufzeit ein enger Austausch der Projektpartner stattfand.

Die Personalfindung und Besetzung der aus Projektmitteln finanzierten Personalstellen beanspruchte teilweise mehr Zeit als ursprünglich geplant. Zudem kam es an einem

Partnerinstitut (TIB) im späteren Projektverlauf zu mehreren Verzögerungen durch Personalwechsel. Auch die Corona-Pandemie und die damit einhergehenden Einschränkungen des persönlichen Austausches hatten Einschnitte zu Beginn des Projektes zur Folge, insbesondere beim Community Engagement. Daher wurde die Projektlaufzeit bis zum 31.08.2023 kostenneutral verlängert, was zu einem erfolgreichen Abschluss der Aktivitäten führte.

3. Ergebnisse und Erfolge

Die Ergebnisse des DiASPora-Projektes wurden bisher in drei Publikationen veröffentlicht – zwei davon in begutachteten, internationalen wissenschaftlichen Zeitschriften. Drei weitere Manuskripte sind zurzeit in einem fortgeschrittenen Stadium und sollen noch im Jahr 2024 eingereicht und veröffentlicht werden. Darüber hinaus wurden die DiASPora-Ergebnisse in 5 Vorträgen und 6 Postern auf wissenschaftlichen Tagungen vorgestellt.

Die Ergebnisse des Projektes konnten des Weiteren wie geplant in der Bac*Dive*-Datenbank veröffentlicht werden. Dazu zählen insbesondere die durch Text Mining beschleunigten Annotationen der Primärliteratur, die zu Bac*Dive*-Stämmen assoziierten Genomdaten, sowie die KI-gestützten Vorhersagen phänotypischer Eigenschaften. Die Datensätze wurden dabei entsprechend markiert und das DiASPora-Projekt als Referenz angegeben. Der erstellte Knowledge Graph wird in verschiedenen Initiativen der Partnerinstitutionen (u.a. NFDI4Microbiota) nachgenutzt und eingebunden werden.

Zusätzlich wurde mit der Media*Dive* Datenbank von Kulturmedien und Wachstumsbedingungen im Rahmen des Projektes eine komplett neue standardisierte Ressource geschaffen, die in den vergangenen 12 Monaten durchschnittlich etwa 5.000 Nutzer:innen pro Monat verzeichnen konnte. Außerdem wurden innerhalb der Projektlaufzeit zwei Online-Workshops zu Bac*Dive* und Media*Dive* gegeben, die beide der Öffentlichkeit offenstanden und zwischen 30 und 50 teilnehmende Personen verzeichnen konnten.

Die Erfolge des Projektes wurden ebenfalls über Pressemitteilungen kommuniziert. So gab es am 18.12.2019 eine gemeinsame Pressemitteilung aller Projektinstitutionen zum Start des Projektes: "Mikrobiologie trifft auf Informatik - Forschungsprojekt DiASPora zur Biodiversität von Bakterien erfolgreich im Leibniz-Wettbewerb". Eine weitere Pressemitteilung vom 02.11.2022 verlautete “Media*Dive*: Rezeptdatenbank unterstützt die globale Erforschung der Biodiversität” und resultierte in den folgenden Presseveröffentlichungen:

- Datenbank für die Anzucht von Mikroorganismen (labo.de vom 09.11.2022)
- Umfassender Medienrezepte-Katalog für Mikroorganismen (biooekonomie.de vom 09.11.2022)
- Auf der Spur der Mikroorganismen - Neue Datenbank ist online (regionalheute.de vom 05.11.2022)
- Rezeptdatenbank unterstützt die globale Erforschung der Biodiversität (dielinde.online vom 03.11.2022)
- Neue Datenbank für die Anzucht von Mikroorganismen (analytik.news vom 03.11.2022)
- Neue Datenbank für Anzuchtmedien (leibniz-gemeinschaft.de vom 02.11.2022)
- Neue Datenbank: Anzuchtmedien für mehr als 44.000 Bakterien und Pilze (laborpraxis.vogel.de vom 02.11.2022)

4. Chancengleichheit, Karriereförderung und Internationalisierung

Entsprechend der Gleichstellungsstandards der Leibniz-Gemeinschaft wurde stets darauf geachtet, Diversität zu fördern und allen Menschen prinzipiell gleiche Chancen zu gewähren, einschließlich z.B. Frauen, Eltern, Minderheiten und Schwerbehinderten. In Abstimmung mit den Gleichstellungsbeauftragten der beteiligten Institutionen wurden bei Einstellungen von Personal Frauen, Minderheiten und Schwerbehinderte bei gleicher fachlicher Eignung bevorzugt. Von den insgesamt vier eingestellten Personen waren zwei männlich (Halder, Shahi) und zwei weiblich (Koblitz, Greco). An der DSMZ wird das familiengerechte Personalmanagement durch das Audit *Beruf und Familie* überprüft und zertifiziert (www.berufundfamilie.de). ZB MED hat das *Total E-Quality* Prädikat und die Charta der Vielfalt unterschrieben. Bei gleicher Eignung werden Frauen und Personen mit Schwerbehinderung in der Einstellung bevorzugt. Ebenso achtet die TIB im Rahmen des Prädikats *Total E-Quality* auf ausgezeichnete Bedingungen für die Vereinbarkeit von Beruf und Familie für alle.

5. Strukturen und Kooperationen

Die Kooperation zwischen den Beteiligten erfolgte wie geplant und im Antrag dargestellt.

6. Qualitätssicherung

An allen beteiligten Instituten gelten die Regeln zur Sicherung guter wissenschaftlicher Praxis entsprechend der Leitlinie der Leibniz-Gemeinschaft und/oder dem Kodex der Deutschen Forschungsgemeinschaft. Weitere Einzelheiten wurden in einem Kooperationsvertrag zwischen den beteiligten Institutionen geregelt.

Alle an DiASPora beteiligten Forschenden waren in ständigem Austausch über den Fortgang des Projekts und gemeinsame Publikationen wurden rechtzeitig untereinander abgestimmt. Die wissenschaftlichen Publikationen aus DiASPora erfolgten in sorgfältig ausgewählten, begutachteten, internationalen Fachzeitschriften (peer-review). Diese Publikationen sind frei verfügbar (Open Access). Die im Projekt generierte Software wurde guter Softwareentwicklungspraxis folgend entwickelt und unter OSI-konformen Open Source Lizenzen veröffentlicht.

7. Zusätzliche Ressourcen

Die folgenden 'in-kind'-Leistungen wurden im Rahmen von DiASPora erbracht:

DSMZ:
Kuratoren und Kuratorinnen der DSMZ haben aktiv an der Entwicklung und Kuration der Kulturmedien in Media*Dive* mitgearbeitet und die vorhergesagten Kulturmedien getestet. Darüber hinaus haben studentische Hilfskräfte bei der Annotation von Literaturdaten und der Überprüfung der Ergebnisse der Annotationsplattform unterstützt. Die Personalmittel werden auf 4 Personenmonate pro Jahr geschätzt.

TIB:
Kuratoren der TIB haben im Rahmen des Projekts eine sehr umfangreiche Ontologie für die Bac*Dive*-Datenbank entwickelt. Dabei wurden mehrere Wissensgraphen für die Datenbank generiert, da diese sich im Projektverlauf weiterentwickelt hat. Zusätzlich wurden mehrere

etablierte Senior-Wissenschaftler:innen kontaktiert, um die Konzepte mit den anderen Partnerinstitutionen sowie mit bekannten veröffentlichten Ontologien zu harmonisieren. Die Personalmittel werden auf 3 Personenmonate pro Jahr geschätzt.

ZB MED:
Die Arbeit von Herrn Halder wurde durch einen Webentwickler von ZB MED mit etwa 5 Personenmonaten unterstützt.

8. Ausblick

Die vielversprechenden Ergebnisse des Projektes werden innerhalb der DSMZ aufgegriffen und dienen als Basis für weitere Vorhaben, die nun aus institutionellen Mitteln gefördert werden. Der Knowledge-Graph soll dabei auf alle Datenbanken der DSMZ ausgeweitet werden, die Text Mining-gestützte Annotationsplattform soll für die BRENDA Enzymdatenbank angepasst werden und die KI-gestützten Vorhersagen auf Basis von Genomsequenzen sollen um weitere Eigenschaften und Organismengruppen erweitert werden. Außerdem wird der erstellte Knowledge Graph in weiteren Initiativen, wie beispielsweise NFDI4Microbiota, nachgenutzt werden.

Die TIB hat einen visuellen Erkundungsmechanismus in die KG integriert, der es den Nutzer:innen ermöglicht, nicht nur SPARQL-Abfragen zu stellen und tabellarische Daten abzurufen, sondern auch die Ergebnisse als Graphen zu sehen. Diese Graphen sind interaktiv, so dass die Nutzenden beispielsweise hinein- und herauszoomen, die Daten auf einem bestimmten Pfad durchlaufen und mehr Informationen über die Beziehungen zwischen verschiedenen Datenelementen erhalten können. Dies ist ein wesentliches Ergebnis des Projekts und wird dazu beitragen, die Beziehungen zwischen den Datensätzen nicht nur in diesem Projekt, sondern auch in künftigen mikrobiellen Studien besser zu verstehen.