GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Schlussbericht

horizontal4meta – Ein neues Paradigma der Sequenzsuche zur Nutzung der metagenomischen Datenflut

Förderkennzeichen: 031L0185
CompLS – Runde 1

Projektleitung:
Dr. Johannes Söding
Quantitative Biologie und Bioinformatik
Max-Planck-Institut für Multidisziplinäre Naturwissenschaften
Am Faßberg 11
37077 Göttingen
soeding@mpinat.mpg.de

Laufzeit des Vorhabens: 01.04.2019 – 31.12.2023

Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Teil I: Kurzbericht

1. Ursprüngliche Aufgabenstellung

In der Shotgun-Metagenomik werden die gemomische DNA und RNA von Bakterien, Archäen und Virusen direkt sequenziert. Die Metagenomic ermöglicht die Analyse der 99% Bakterien und Archäen, die bisher nicht im Labor kultivierbar sind. Dadurch revolutioniert sie derzeit die mikrobielle Ökologie und auch die Medizin, da sie zum Beispiel die detaillierte Untersuchung des Mikrobioms im menschlichen Darm erlaubt und im letzten Jahrzehnt der enorme Einfluss unseres Darmmikrobioms auf unser Immunsystem, unser zentrales Nervensystem, unseren Metabolismus und insgesamt unsere Gesundheit deutlich geworden ist. Die Funktion der Proteine in metagenomischen Sequenzen wird üblicherweise durch Annotationstransfer von verwandten, funktionell annotierten Proteinen in Referenzdatenbanken vorhergesagt (”vertikaler Informationstransfer”). Ein Haupthindernis besteht darin, dass oft nur ein Teil der Proteinsequenzen auf diese Weise annotiert werden kann. Selbst die Funktion der annotierten Proteine ergibt sich oft nicht klar aus ihren Annotationen, da der biologische Wirkzusammenhang fehlt.

Ziel dieses Projekts war die Entwicklung eines Hochdurchsatz-fähigen Ansatzes der Proteinfunktionsvorhersage basierend auf ”horizontalem Informationstransfer”. Dieser beruht auf folgender Eigenschaft: Bei Prokaryoten und Viren bevorzugt die Evolution, Gene, die an denselben biologischen Stoffwechselwegen beteiligt sind, als konservierte Gencluster zusammenzuhalten. Umgekehrt ist die Erhaltung der Gen-Nachbarschaft ein starkes Indiz für eine funktionelle Assoziation der Gene. Durch systematische Identifikation zwischen zwei Genomen konservierter Gencluster können wir daher die funktionale Assoziation von Genen der kodierten Proteine vorhersagen. Diese Idee ist zwar seit über 25 Jahren bekannt, allerdings hat bis heute keine Software diese horizontale Information systematisch und automatisiert vermocht zu nutzen. Ein Grund ist die Schwierigkeit, die schnell steigende Anzahl bekannter Genome zu analysieren, da die erforderliche Rechenzeit quadratisch mit der Anzahl der zu analysierender Genome ansteigt. Ein weiterer Grund ist die limitierte Sensitivität der verwendeten Suchmethoden zur Detektion entfernt verwandter Proteine. Ein dritter Grund ist die meist recht unflexible Charakterisierung der konservierten Gencluster als syntenische oder kollineare Sequenzblöcke.

2. Ablauf des Vorhabens

In dem Vorhaben wurden folgende Schritte bearbeitet:

1. Die Entwicklung eines Softwaretools für die taxonomische Annotation von Genen, das die Nachbarschaft der auf dem gleichen Contig (aus Short Reads assemblierten Sequenzen) befindlichen Gene für die taxonomische Zuordnung mit einbezieht.

2. Die Entwicklung der multi-match P-Wert-Statistik, die P-Werte mehrerer Sequenz-Matches zwischen einer Suchsequenzmenge und einer Zielsequenzmenge

kombiniert. Als Anwendung dieser Statistik entwickelten wir das Softwaretool SpacePharer, ein Tool für die *de-novo*-Vorhersage von Phagen-Wirt-Beziehungen durch Identifizierung von Phagengenomen, die mit CRISPR-Spacern in genomischen oder metagenomischen Daten übereinstimmen.

3. Die Entwicklung der Match-Clustering-P-Wert-Statistik, die bewertet, ob der Clusterungsgrad der Sequenzmatches innerhalb des Zielgenoms durch Zufall erklärbar ist oder ob er auf ein wegen funktionaler Assoziation erklärbares, konserviertes Gencluster hinweist. Zusätzlich wurde eine Statistik entwickelt, mit der bewertet werden kann, ob der beobachtete Grad der Kodirektionalität und Genordnung der Clustermatches innerhalb des Zielgenoms durch Zufall erklärbar ist oder ob er auf eine wegen funktionaler Assoziation konservierte Region hinweist.

4. Die Entwicklung des Suchtools Spacedust für konservierte Genclusters, die ähnliche Genclusters zwischen zwei Genomen identifiziert, mit Hilfe der oben erwähnten Statistiken bewertet, die Cluster optimiert und nach Signifikanz ordnet.

5. Die Entwicklung des Softwaretools UniOP zur globalen Vorhersage bakterieller und archäaler Operons.

6. Die globale Vorhersage aller funktionalen Assoziationen zwischen Proteinen in 1308 bakteriellen Referenzgenomen und Annotation mittels vertikalen und horizontalen Informationstransfers.

7. Die Zurverfügungstellung gut dokumentierter open-source-Software-Pakete für Spacedust und UniOP sowie einer Datenbank mit funktionelle Annotationen der analysierten 1308 Referenzgenome.

3. Wesentliche Ergebnisse

Entwicklung des Softwaretools mmseqs taxonomy: github.com/soedinglab/MMseqs2
Zugehörige Publikation: Mirdita M, Steinegger M, Breitwieser F, Söding J, Levy Karin E. Fast and sensitive taxonomic assignment to metagenomic contigs. Bioinformatics. 2021; 37(18):3029-31. Zitiert: 135-mal.

Entwicklung des Softwaretools „SpacePharer“: github.com/soedinglab/spacepharer
Zugehörige Publikation: Zhang R, Mirdita M, Levy Karin E, Norroy C, Galiez C, Söding J. SpacePHARER: sensitive identification of phages from CRISPR spacers in prokaryotic hosts. Bioinformatics. 2021; 37(19):3364-6. Zitiert: 57-mal.

Entwicklung eines Softwaretools „UniOP“: github.com/soedinglab/uniop
Zugehörige Publikation: Su H, Zhang R, Söding J. UniOP: A universal operon prediction for high-throughput prokaryotic (meta-)genomic data. 2024; In preparation.

Entwicklung eines Softwaretools „Spacedust“: github.com/soedinglab/spacedust
Zugehörige Publikation: Zhang R, Mirdita M, Söding J. De novo discovery of conserved gene clusters in microbial genomes with Spacedust. 2024. In preparation.

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Schlussbericht

horizontal4meta – Ein neues Paradigma der Sequenzsuche zur Nutzung der metagenomischen Datenflut

Förderkennzeichen: 031L0185
CompLS – Runde 1

Projektleitung:
Dr. Johannes Söding
Quantitative Biologie und Bioinformatik
Max-Planck-Institut für Multidisziplinäre Naturwissenschaften
Am Faßberg 11
37077 Göttingen
soeding@mpinat.mpg.de

Laufzeit des Vorhabens: 01.04.2019 – 31.12.2023

Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Teil II: Eingehende Darstellung der Arbeiten

1. Verwendung der Zuwendung

Der Großteil der bereitgestellten Mittel (86%, 465 k€) wurde für Personalausgaben verwendet, im wesentlichen Teil für zwei PhD-Studenten. 5% (29.636 €) der Mittel wurden für einen High-Memory-Server ausgegeben, der auch für die Weiterführung der hier beschriebenen Arbeiten verwendet wird, um die genomischen und metagenomischen Daten zu halten. Wegen der COVID-19-Pandemie sind kaum Reisen getätigt worden, was zu geringen Kosten für Dienstreisen geführt hat.

Wegen des frühzeitigen Ausscheidens eines Mitarbeiters aus familiären Gründen (M. Mirdita) wurde das Projekt kostenneutral von Ende Februar auf Ende Dezember 2023 verlängert und durch Dr. R. Zhang zuendegeführt. Dadurch konnten alle gesetzten Projektziele erreicht werden.

2. Erzielte Ergebnisse

Die folgenden Abschnitte beschreiben die erzielten Ergebnisse in der Reihenfolge, in der sie in der Grant Application unter Abschnitt 1.3 „Wissenschaftliche und technische Arbeitsziele“ genannt sind.

2.1. Entwicklung eines Softwaretools für die taxonomische Annotation von Genen, welches die Nachbarschaft der auf dem gleichen Contig befindlichen Gene für die taxonomische Zuordnung mit einbezieht

Zusätzlich zu den im Arbeitsplan des Antrags aufgeführten Arbeiten entwickelten wir mmseqs taxonomy (Fig. 1), eine neue Anwendung für die Contig-basierte taxonomische Annotation von Genen. Die Möglichkeit, hier eine wichtige Lücke bei Sequenzannotationsmethoden relativ leicht schließen zu können und die Komplementarität der in horizontal4meta zu entwickelnden Methoden hatte uns veranlasst, diese Arbeiten mit in unser Projekt aufzunehmen. Um den zusätzlichen Personalbedarf für diese zusätzliche Arbeiten zu minimieren, wurden wir in diesem Projekt von einer nicht von horizontal4meta finanzierten postdoktoralen Wisschaftlerin in der Arbeitsgruppe (Dr. Eli Levin Karin) unterstützt.

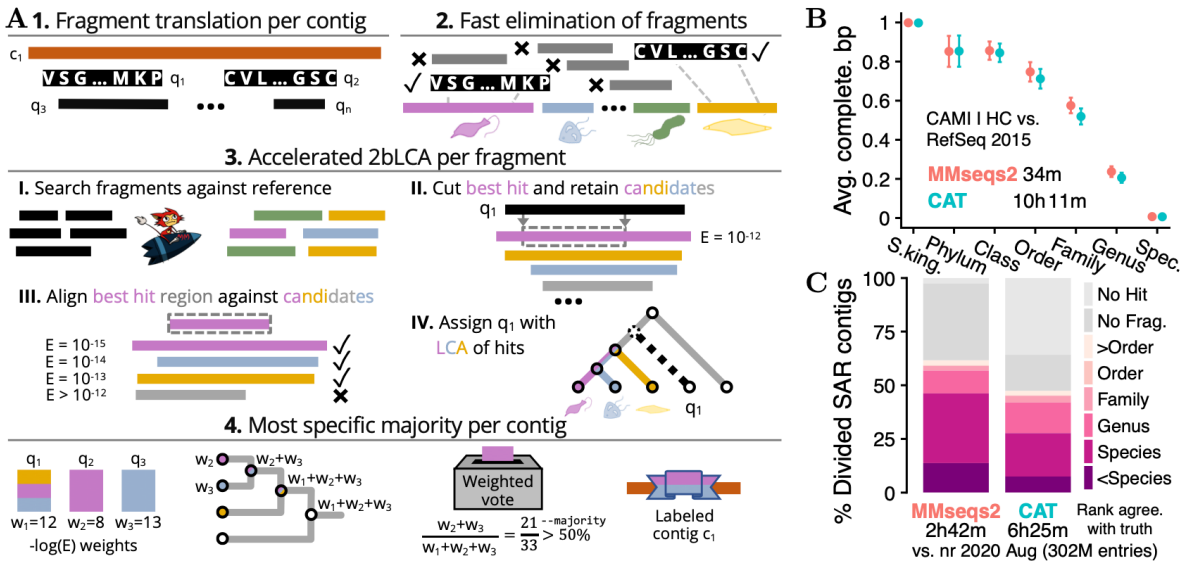

FIG. 1. (**A**) Taxonomy assignment algorithm in four steps: (1) Translate all possible protein fragments in six frames from all contigs. (2) Reject fragments unlikely to find a taxonomic hit in later stages (full details in Supp. Information). (3) Assign taxonomic nodes using an accelerated 2bLCA procedure. Each query fragment q is searched against the reference database, resulting in a list l of all its homologous targets. The aligned region between q and the best hit t (with E-value $E(q,t)$) is aligned against all targets in l. We assign q the LCA of the taxonomic lables of all target sequences that have an E-values lower than $E(q,t)$. By realigning l we could avoid the costly second search of 2bLCA. (4) Each assigned q contributes its weight $(-\log E(q,t))$ to its taxonomic label and all labels above it, up to the root. The contig's taxonomic node is determined as the most specific taxonomic label, which has a support of at least the `--majority` parameter. The support of a label is the sum of its contributing weights divided by the total sum of weights. (**B**) MMseqs2 taxonomy (red) is ∼18x faster and achieves similar average completeness to CAT (turquoise) on a bacterial benchmark. (**C**) Fragments extracted and retained by MMseqs2 from eukaryotic SAR contigs result in more correctly classified contigs than those extracted for CAT by Prodigal, at approximatly twice the speed. Runtimes measured on a 2x14-core Intel E5-2680v4 server with 768GB RAM.

2.2. Entwicklung der multi-match-P-Wert-Statistik, die P-Werte mehrerer Sequenz-Matches kombiniert, und Anwendung im Softwaretool SpacePharer

Im Antrag schlugen wir den ersten allgemeinen statistischen Ansatz zur Berechnung exakter P-Werte zur Bewertung von Mehrfach-Sequenzmatches vor. Dieser Ansatz ermöglicht es uns, verwandte Sequenzgruppen zu erkennen, selbst wenn alle paarweisen Sequenzmatches für sich genommen weit davon entfernt sind, statistische Signifikanz zu erreichen. Die im Antrag dargelegte Berechnung der Multi-match-P-Wert-Statistik wurde implementiert und erfolgreich getestet. Die Multi-match-P-Wert-Statistik ist Basis für das Auffinden der Clustermatches in unserer Spacedust-Software (siehe Abschnitt 2.4).

Als einfache Anwendung der Statistik entwickelten wir das Softwaretool SpacePharer. Dieses Tool erlaubt es für Prokaryonten mit gegebenem Genom inklusive CRISPR-Arrays vorherzusagen, welche Phagen den Prokaryonten befallen können. Dazu werden die etwa 30 Nukleotide langen Spacer Sequenzen in den CRISPR-Arrays des Wirtes gesucht gegen alle bekannten Phagengenome. Wird ein signifikanter Einzel- oder Mehrfach-Sequenzmatch gefunden, so ist der Nachweis der Infektion des Wirtes mit dem entsprechenden Phagen erbracht. Da die Spacer-Sequenzen sehr kurz und damit wenig informativ sind, ist die Verwendung von Mehrfachmatches sehr hilfreich. Zur statistischen Bewertung dieser Mehrfachmatches wiederum war unsere neue Multi-match-Statistik wichtig. Die Performance von SpacePharer erwies sich als dem Standardtool für diese Aufgabe, BLASTN, überlegen (Fig. 2).

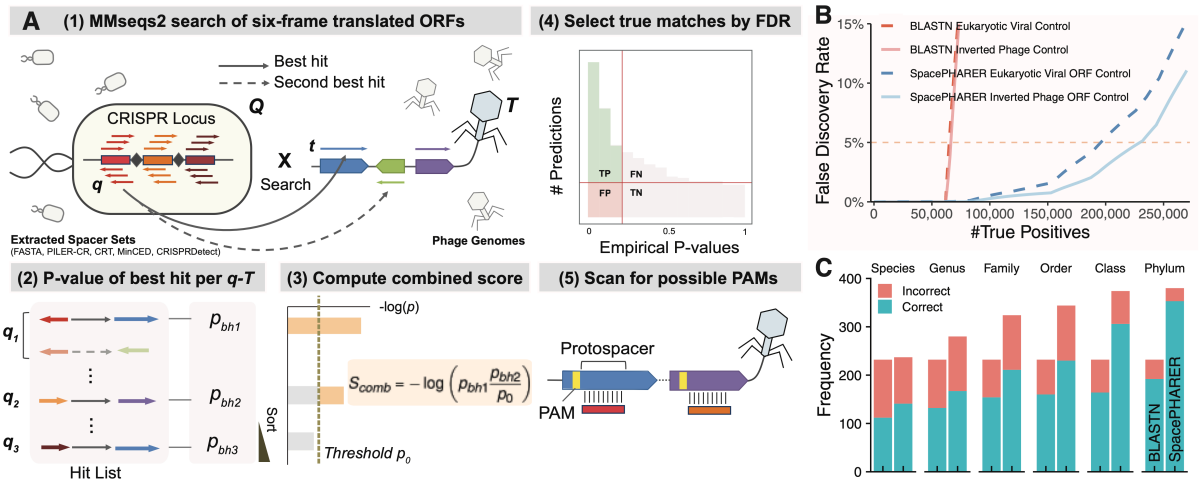

Abb. 2: (**A**) SpacePharer-Algorithmus. Ein Abfragesatz Q besteht aus übersetzten 6-Rahmen-ORFs (q) von CRISPR-Spacern, und ein Zielsatz T besteht aus übersetzten 6-Rahmen-ORFs (t) von Phagenproteinen. (1) Suche aller q's gegen alle t's. Alignieren der q-t-Treffer auf Nukleotidebene und Priorisieren der Nukleotidtreffer. (2) Für jedes q-t-Paar: berechnen des P-Werts für den besten Treffer aus der Statistik erster Ordnung. (3) Berechnen des S_{comb}-Scores, indem die P-Werte des besten Treffers aus mehreren Treffern zwischen Q und T kombiniert werden. (4) Schätzen der False Discovery Rate (FDRs). (5) Suche nach möglichen Protospacer-Assoziierten Motiven (PAM). (**B**) Vergleich zwischen SpacePharer (blau) und BLASTN (rot) unter Verwendung von invertierten Phagensequenzen (durchgezogene Linien) oder eukaryotischen viralen ORFs als Null-Set (gestrichelte Linien), dargestellt anhand der erwarteten Anzahl von true positive (TP) Vorhersagen bei unterschiedlichen FDRs. (**C**) Vergleich zwischen BLASTN (links) und SpacePharer unter Verwendung des gewichteten Verfahrens des niedrigsten gemeinsamen Vorfahren (LCA, rechts) bei FDR = 0,02, ausgewertet anhand der Anzahl richtiger (blau) und falscher (rot) Vorhersagen für alle Wirtsvorhersagen, die auf jedem taxonomischen Rang oder darunter gemacht wurden.

2.3. Entwicklung der Match-Clustering- und der Kodirektionalität-Statistik zur Bewertung der Signifikanz von Clusterähnlichkeit

Die Konservierung der Gennachbarschaft ist eine Signatur funktionell assoziierter Gene, da ein Auseinanderreißen funktional assoziierter Gene tendenziell die Fitness der betroffenen Nachkommen beeinträchtigt. Um konservierte Gencluster aufzuspüren, benötigen wir ein Maß für die Konservierung von Cluster-Matches. Die Konservierung kann nach zwei Kriterien bewertet werden.

Erstens: Wie wenig zufällig ist die Anreicherung von Sequenzmatches in engen Bereichen des Such- sowie des Zielgenoms? Intuitiv ist klar, dass die Signifikanz eines Cluster Matches desto höher ist, je mehr Gene Proteine zwischen dem Bereich in Suchgenom und dem Bereich im Zielgenom einander zugeordnet sind, und je weniger nicht zugeordnete Gene in den beiden Bereichen liegen. Der Grad der Zufälligkeit kann quantifiziert werden als Wahrscheinlichkeit, dass *k* oder mehr Gene zugeordnet wurden in einem Bereich von jeweils *m* Genen im Such- und Zielgenom unter der Annahme, dass die Sequenzmatches zwischen beiden Genomen völlig zufällig verteilt

sind (Nullmodell). Im Antrag hatten wir hierzu einen Vorschlag gemacht, den wir im Vorhaben überarbeitet und verbessert haben, um eine schnellere Berechnung zu ermöglichen. Der neue Clustering-P-Wert ist $m!^2\ q_0^k\ /\ (m-k)!^2\ /\ k!$, wobei q_0 die Wahrscheinlichkeit ist, dass zwischen einem beliebigen Genpaar aus Such- und Zielgenom ein Match gefunden wird. Die mathematische Ableitung ist im Supplemental Material des Spacedust-Manuskriptes enthalten.

Zweites Kriterium ist: Wie wenig zufällig ist die Ähnlichkeit der Reihenfolge und des Stranges der Proteine im Suchgenom zu den ihnen zugeordneten Proteinen im Zielgenom? Intuitiv ist klar, dass die Signifikanz eines Cluster Matches desto höher ist, je mehr aufeinanderfolgende Paare im Suchgenom auch im Zielgenom aufeinanderfolgen und auf den gleichen Strängen kodiert sind. Der Grad der Zufälligkeit kann quantifiziert werden als Wahrscheinlichkeit, dass *m* oder mehr Paare aufeinanderfolgender Gene im Suchgenom auch im Zielgenom aufeinander folgen und auf dem gleichen Strang kodiert sind, unter der Annahme, dass die Sequenzmatches zwischen beiden Genomen völlig zufällig verteilt sind (Nullmodell). Im Antrag hatten wir hierzu einen Vorschlag gemacht, der eine approximative Berechnung erlaubte. Diesen haben wir im Vorhaben überarbeitet und die neuen Lösung ist exakt und sehr effizient berechenbar: Der neue Kodirektionalitäts-P-Wert ist einfach $(1-n/k)/(2^n\ n!)$. Die mathematische Ableitung ist im Supplemental Material des Spacedust-Manuskriptes enthalten.

Beide Statistiken wurden mit synthetischen Daten evaluiert und haben sich als genau und robust erwiesen. Beide Statistiken sind die Basis für das Auffinden und Optimieren der Clustermatches in unserer Spacedust-Software (siehe nächster Abschnitt).

2.4. Entwicklung eines Tools Spacedust zur Suche konservierter Gencluster, welches die Ähnlichkeit der Gencluster mit Hilfe der im vorigen Abschnitt beschriebenen Statistiken bewertet und optimiert

Ein Hauptziel dieses Vorhabens war die Entwicklung von Spacedust, einer schnellen und sensitive Software für die systematische *de-novo* Entdeckung konservierter Gencluster in den Eingabe-Genomen (Abbildung 3). Hierzu vergleicht es alle Proteine der Suchgenome mit allen Proteinen der Zielgenome (wobei beide Mengen identisch sein dürfen: „alle-gegen-alle-Analyse“). Für maximale Sensitivität zur Detektion entfernt homologer Proteine nutzt es den schnellen Strukturvergleich mit unserer Foldseek Software (Steinegger 2022). Für jedes Paar von Such- und Zielgenom wird eine Liste mit Protein-Protein-Hits ähnlicher Struktur erstellt und die Hits auf räumliche Clusterung im Such- und Zielgenom hin untersucht. Beginnend mit allen einzelnen Protein-Protein-Hits als kleinster möglicher Cluster-Match wird jedes Cluster um mögliche weitere Hits erweitert, falls die statistische Signifikanz dadurch erhöht wird, bis kein Cluster-Match durch Erweiterung mehr an Signifikanz gewinnen kann. Die so erhaltenen, maximal signifikanten Cluster-Matches werden dann als Suchresultate ausgegeben.

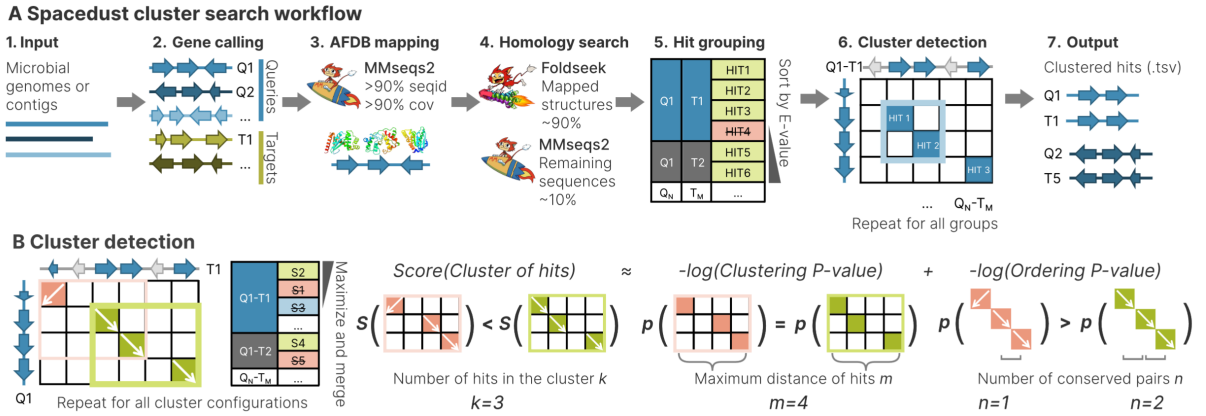

Abb. 3: Spacedust-Algorithmus. (**A**) 1. Eingabe sind mikrobielle Suchgenome (Q1,...) und Zielgenome (T1,...). 2. Vorhersage Protein-kodierender Gene in allen Genomen. 3. Abruf der Proteinstrukturen aller kodierter Proteine aus der AlphaFold DB (AFDB) mit MMseqs2. 4. Suche aller Suchproteine gegen alle Zielproteine mit Foldseek und MMseqs2 und kombinieren der Suchergebnisse. 5. Gruppierung der Treffer nach Q-T-Paaren, Auswahl des besten Treffers für jede Abfragesequenz in T und Auswahl der besten Treffer nach ihrem E-Wert. 6. Für jedes Q-T-Paar: Ermittlung der Cluster-Matches mit signifikant konservierter Gen-Nachbarschaft. 7. Ausgabe der paarweisen Cluster-Matches zwischen Q-T in einer Tab-separierten Datei. (**B**) Ausgehend von Clustern bestehend aus einem einzigen Protein-Protein-Match werden die Cluster iterativ erweitert, um die statistische Signifikanz zu maximieren. Dabei wird die Signifikanz der Konservierung der Gen-Nachbarschaft durch die Summe der Logarithmen des Clustering-P-Werts und des Ordnungs-P-Werts bewertet.

Spacedust ist das erste Softwaretool zur *de-novo*-Detektion konservierter Gencluster, das eine alle-gegen-alle-Analyse einer großen Anzahl von Genomen ermöglicht. Es ist bisherigen Methoden in mehrfacher Hinsicht überlegen: (i) Es benötigt keine Referenz-Datenbank und kann daher konservierte Cluster jeder Art und Zusammensetzung *de-novo* entdecken. (ii) Die strukturbasierte Suche steigert die Empfindlichkeit der Detektion sehr entfernt verwandter Proteine und damit auch sehr entfernt verwandter konservierter Gencluster. (iii) Die hohe Geschwindigkeit ermöglicht die Suche nach konservierten Clustern in einer großen Anzahl von Genomen. (iv) Spacedust integriert die funktionale Annotation der Gene/Proteine mit Hilfe von eggNOG (Huerta-Cepas 2015). (v) Wir stellen ein benutzerfreundliches Google Colab Notebook für Spacedust zur Verfügung.

Die geringe Geschwindigkeit bisheriger Softwaretools zum Aufspüren konservierter Gencluster wie Gecko3 (Winter 2013) und OrthoGNC (Jahangiri-Tazehkand, 2017) erlaubte uns nicht deren Anwendung und Test auf größeren Genom-Datensets. Um Spacedusts Leistungsfähigkeit zu testen, haben wir es daher mit einem auf eine spezifische Klasse von Genclustern spezialisierten Tools verglichen: Antiviral Defense-Systeme (AVS) und biosynthetische Gen-Cluster (BGCs).

PADLOC ist die populärste Software zur Detektion antiviraler Abwehrsysteme, die auf einer großen Referenzdatenbank mit Profil-HMMs beruht, die alle bekannten, an

antiviralen Defense-Systemen beteiligten Proteinfamilien kodieren. Da keine Genome mit „Goldstandard-Annotation“ von Antiviralen Abwehrsystemen existiert, konnten wir lediglich vergleichen, welchen Anteil der von PADLOC als Antiviral Defence Islands annotierten Gencluster ebenfalls von unserem viel allgemeineren, referenzlosen De-Novo-Discoverytool Spacedust identifiziert wurden. Wir haben PADLOC sowie Spacedust auf bakteriellen Referenzgenomen laufen lassen. Abbildung 4 zeigt die Ergebnisse dieser Analyse. Spacedust konnte insgesamt 5255 (95%) von den 5520 von PADLOC annotierten antiviralen Defense-System-Clustern aufspüren, wobei alle bis auf 367 der Cluster vollständig mit der mit der PADLOC-Annotation übereinstimmten. Mit anderen Worten: 95% aller in langwieriger Arbeit entdeckten und in diversen Publikationen beschriebenen Abwehrsysteme vermochte Spacedust, ohne jegliche Vorkenntnisse irgendwelcher antiviraler Abwehrsysteme oder derer Proteine, in Referenzgenomen zu identifizieren. Diese erstaunliche Empfindlichkeit illustriert eindrucksvoll die Macht dieser Methode für die Entdeckung bisher gänzlich unbeschriebener Systeme, sei es für antivirale Abwehr, sei es für vielzählige weitere bekannte und unbeschriebene Funktionen.

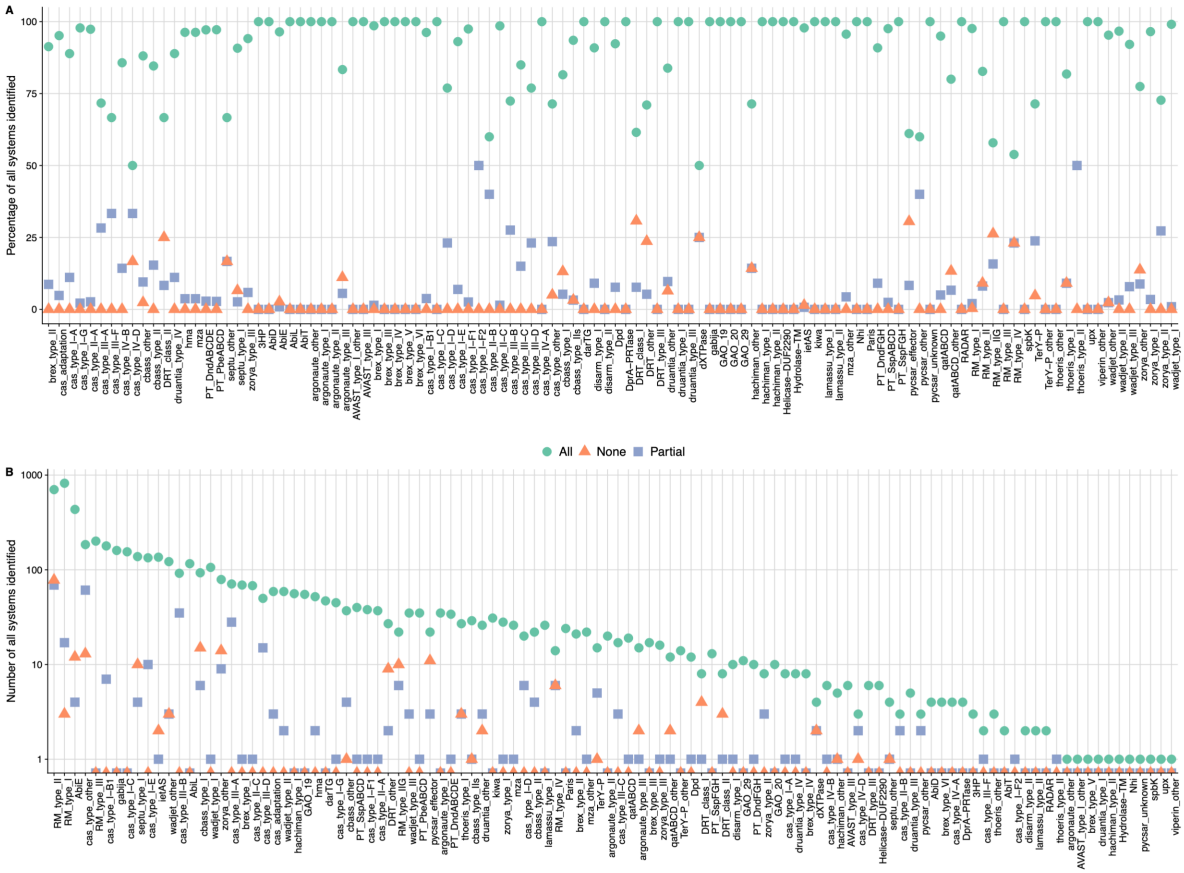

Abb. 4: (**A**) Prozentuale und (**B**) absolute Anzahl der von PADLOC identifizierten antiviralen Abwehrsysteme, die von Spacedust vollständig (grüner Kreis), teilweise (blaues Quadrat) oder nicht (rotes Dreieck) detektiert wurden. Jede Spalte bezeichnet einen Typ von antiviralem Abwehrsystem.

2.5. Globale Vorhersage funktionaler Assoziationen zwischen Proteinen in repräsentativen Genomen

Der Vergleich mit PADLOC und zwei Tools zum Auffinden von biosynthetischen Gen-Clustern demonstriert die Empfindlichkeit von Spacedust für das Aufspüren konservierter, funktional assoziierter Gencluster. Für eine globale Analyse haben wir ein bakterielle Referenzdatenbank erstellt. Die 7167 vollständigen bakteriellen Genome hoher Qualität aus der KEGG GENOME-Datenbank haben wir mit 70% maximaler paarweiser durchschnittlicher Nukleotididentität gefiltert, was etwa der Genus-Ebene entspricht, und so 1308 repräsentativen Genome mit insgesamt 4,2 Millionen kodierten Proteinen erhalten.

Zur systematischen Identifikation sämtlicher konservierter Gencluster berechneten wir dann mit Spacedust die zwischen jedem der 854 778 Paare von Genomen die konservierte Clustermatches. Dies erforderte unter anderem $(4{,}2\times10^{6})^2 = 17{,}6\times10^{12}$ paarweise Proteinstrukturvergleiche erforderte. Dank unserer Foldseek-Software konnte Spacedust diese Strukturvergleiche in nur 72 Stunden auf zwei Servern mit jeweils 256 CPU-Kernen bewältigen konnte, also in etwa 4μs pro paarweisem Vergleich pro CPU-Kern.

Diese Analyse förderte 107 Millionen paarweise konservierte Gencluster (Clustermatches) mit insgesamt 321.2 Proteinen zu Tage. Ein Großteil dieser Gencluster ist nicht als Modul oder Genfamilie beschrieben. Für eine quantitative Bewertung, in wie weit diese Gencluster funktional assoziierten Genmodulen entsprechen, haben wir alle 4,2 Millionen Gene/Protein in den 1308 Referenzgenomen mit der eggNOG-Pipeline (Cantapiedra 2021) funktional annotiert. Unter anderem weist diese Software soweit möglich Identifikatoren (IDs) für Stoffwechsel-Pathways der KEGG-Datenbank (Kanehisa 2023) zu. Dies gestattete uns die Definition von (vermutlich) falschen und richtigen Vorhersagen: Alle Genepaare innerhalb desselben Genoms, die in mindestens *n* Genclustermatches gemeinsam vorkommen und deren Gene beide einen KEGG-Stoffwechsel-Pathway von eggNOG zugeordnet bekommen haben, werden als True Positives (TP) gewertet, wenn beide Pathway-IDs identisch sind und als False Positives (FP), wenn sie sich unterscheiden. Dies erlaubt nun eine Genauigkeits-Empfindlichkeits-Analyse (Precision-Recall). Genauigkeit (Precision) ist der Anteil der korrekten Vorhersagen (TP) geteilt durch die gesamte Anzahl an Vorhersagen (TP+FP). Empfindlichkeit (Recall) ist der Anteil der korrekten Vorhersagen (TP) an der Gesamtzahl möglicher korrekter Vorhersagen. Für alle Werte von *n*=2 bis Unendlich wird die entsprechende Genauigkeit und Empfindlichkeit im Graphen aufgetragen. Die Analyse in Abbildung 5 zeigt, dass Spacedust bei einer hohen Genauigkeit von 90% fast 50% aller möglichen korrekten Vorhersagen tatsächlich vorhersagt. Als Beispiel erhöht sich die Wahrscheinlichkeit, dass zwei Genen mit 3 dazwischenliegenden Genen zum gleichen KEGG-Pathway gehören, durch einen einzigen Clustermatch mit einem der 1307 anderen Genome von 22% auf 49%, durch 5 Clustermatches auf 60%. Bei dieser Analyse muss bedacht werden, dass die hier geschätzte Genauigkeit die tatsächliche Genauigkeit erheblich unterschätzt,

da Proteine funktional assoziiert sein können auch wenn sie unterschiedlichen KEGG-Pathways zugeordnet sind.

Leider konnten wir Spacedust mit keiner anderen Software vergleichen, da keine auch nur entfernt in der Lage war, in angemessener Zeit alle 1308 Genome zu analysieren.

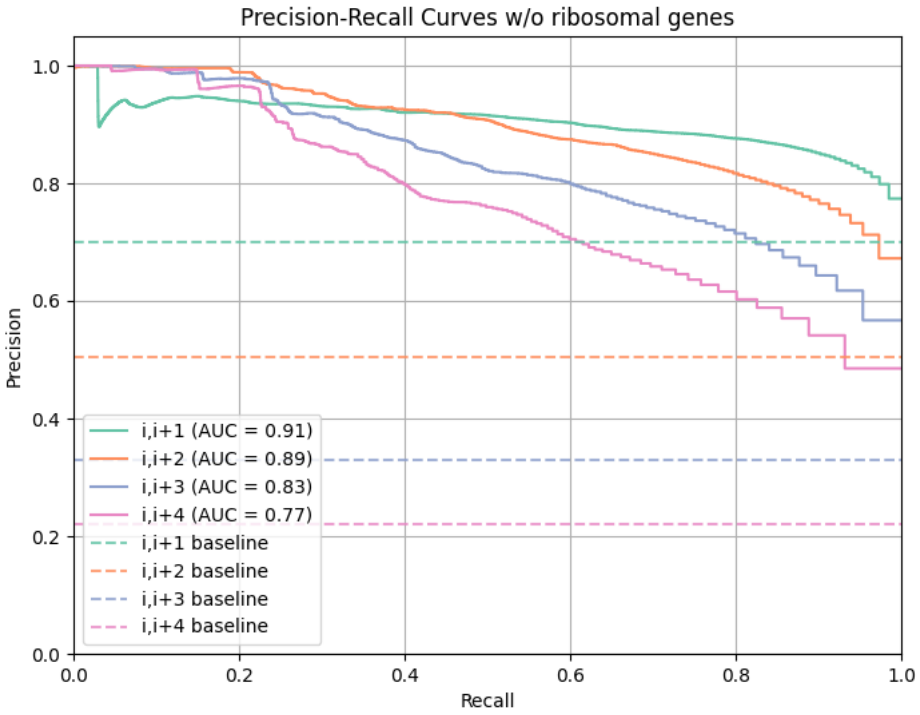

Abb. 5: Genauigkeit (precision) und Empfindlichkeit (recall) von Spacedust, bewertet durch Kongruenz der KEGG-Pathway-IDs der SpaceDust-Cluster-Matches für alle Genpaare, die durch bis zu 4 Gene voneinander getrennt sind. Die gestrichelten horizontalen Linien geben die Genauigkeit ungeachtet von Clustermatches an. AUC = Area Under the Curve.

Um funktionale Assoziation von Genen vorherzusagen, ist neben der Konservierung der Gennachbarschaft die Vorhersage von Operons sehr nützlich. Operons sind Gruppen von auf dem gleichen Strang kodierter Gene, die als Ganzes in einem Transkript abgeschrieben werden, um dann in mehrere, einzelne Proteine übersetzt zu werden. Da die in Operons kodierten Proteine stark ko-reguliert sind (es gibt ja nur einen Promoter für das gesamte Operon), in Operons kodierte Protein in der Regel eng funktional assoziiert. Daher war ein Bestandteil dieses Vorhabens die Entwicklung einer Methode zur Vorhersage von Operons.

Hauptziel bei der Entwicklung unserer UniOP-Methode war, dass keinerlei Informationen außer dem Genom selbst benötigt werden. Insbesondere sollte nicht auf vorhandenen, annotierten Genomen trainiert werden, wie die vielen auf maschinellem Lernen beruhende Methoden dies tun, da bei solchen Methoden mit Overfitting und schlechter Performance auf Genomen gerechnet werden muss, die keinen engen

Verwandten im Trainingsset haben. Einen Überblick über die Funktionsweise der UniOP Methode gibt Abbildung 1.

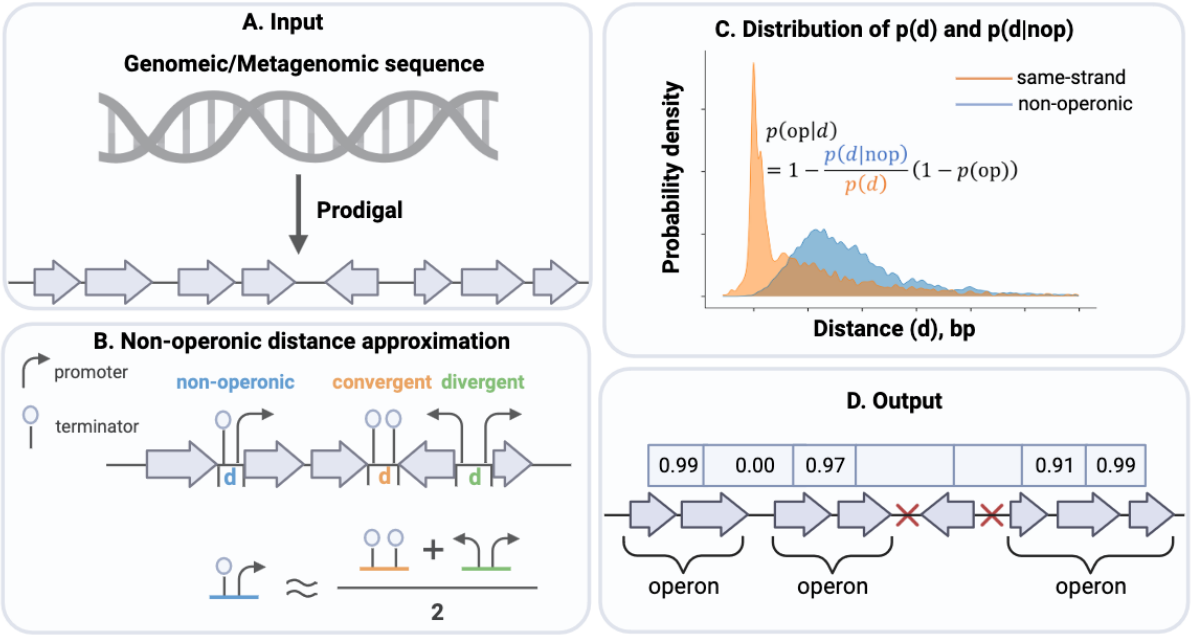

Abb. 6: UniOP-Algorithmus. A. Vorhersage von proteinkodierenden Genen in der eingegebenen genomischen Sequenz. B. Abschätzung der Verteilung der intergenischen Abständen für nicht-operonische Paare. C. Abschätzung der Intergen-Abstandverteilung von gleichsträngigen Genpaaren (p(d)) und von nicht-operonischen Genpaare (p(d|nop)), sowie Abschätzung der Prior Wahrscheinlichkeit p(op) dafür, dass ein beliebiges Genpaar auf dem gleichen Strang operonisch ist. Daraus wird die Wahrscheinlichkeit p(op|d) unter Verwendung des Bayes'schen Theorems vorhergesagt. D. Vorhersagen für jedes benachbarte Paar desselben Strangs und jedes Operons.

Abbildung 7 zeigt Ergebnisse auf 10 Genomen, deren Operons (teilweise) annotiert wurden. UniOP schneidet auf allen Genomen entweder besser ab als das nächstbeste Tool oder ähnlich gut.

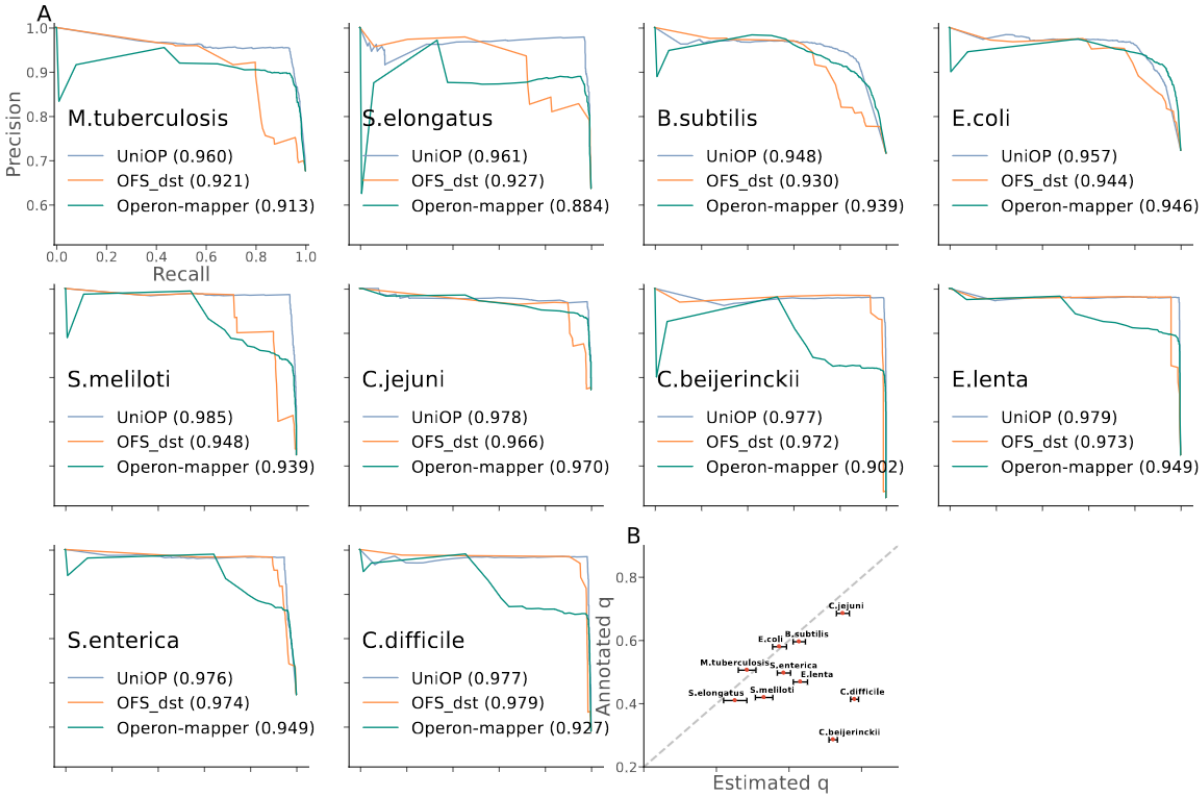

Abb. 7: Vergleich der Genauigkeit von OFS_dst (Westover 2005), Operon-mapper (Taboada 2018) und UniOP auf 10 bakteriellen Genomen mit annotierten Operons. Die Fläche unter den Genauigkeits-Empfindlichkeits-Graphen (Precision-Recall) ist jeweils in Klammern vermerkt. Je höher dieser Wert ist, desto besser sind die Vorhersagen.

2.6. Annotation 1308 bakterieller Referenzgenome mittels vertikalen und horizontalen Informationstransfers

Spacedust stellt für jedes der 1308 Referenzgenome die Clustermatches mit und die identifizierten Gencluster zur Verfügung. Zur Illustration dieses umfänglichen Datensatzes ist in Abbildung 8 ein kleiner Teil dieser Daten, nämlich für Genindizes 500 bis 800 eines einzigen Genoms, des Cyanobakteriums *Synechocystis*, dargestellt. Diese Ansicht entspricht also lediglich 300/4,2E6 = 0,007% des gesamten Datensatzes.

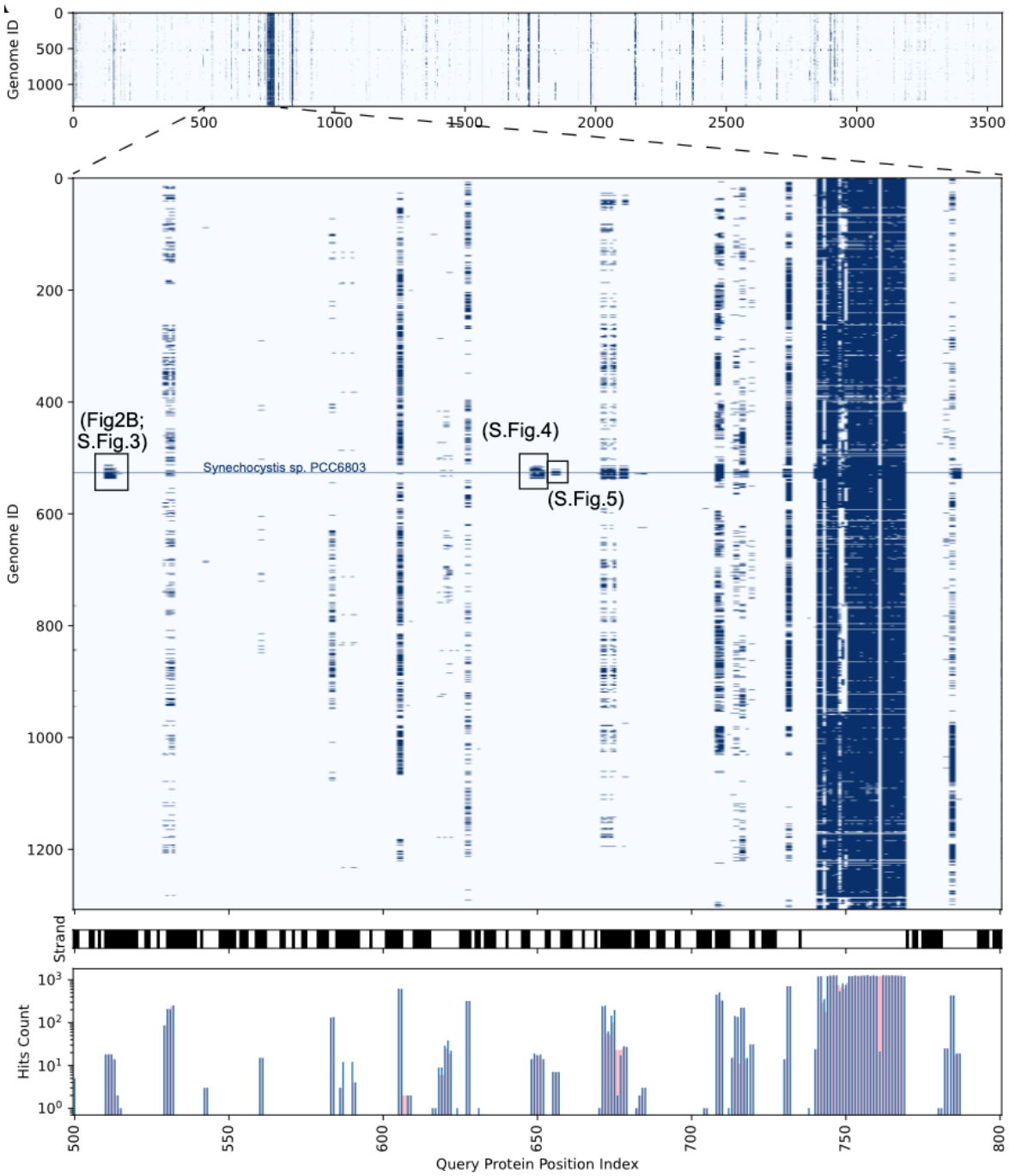

Abb. 8: Durch Spacedust identifizierte konservierte Gencluster am Beispiel des Cyanobakteriums *Synechocystis*, ananlysiert zusammen mit insgesamt 1.308 bakteriellen Referenzgenomen. Oberestes Feld: Cluster-Heatmap des Vorhandenseins bzw. Nichtvorhandenseins von Clustertreffern zwischen Synechocystis mit einem der anderen 1307 Referenzgenome. Jede Zeile zeigt die Clustermatches von Synechocystis mit einem Genom. Mittlere Tafel: Wie obere Tafel, aber für einen Ausschnitt des Genoms zwischen Genpositionen 500 und 800. Die Transkriptionsrichtung der Gene ist darunter in schwarz (vorwärts) und weiß (rückwärts) widergegeben. Unteres Tafel: Anzahl der geclusterten Treffer pro Protein (blau) und Trefferpaare im selben Gencluster (rosa).

Als Beispiel für die Nützlichkeit der enthaltenden Annotationen zeigen Abbildungen 9 und 10 im Genom von *Synechocystis* identifizierte Gencluster und die homologen Bereich in anderen Genomen.

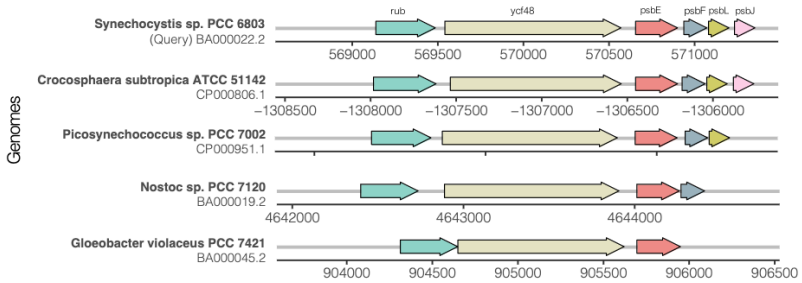

Abb. 9: Durch Spacedust identifizierte Gencluster für das Photosystem II. Orthologe Gene/Proteine sind gleich gefärbt. Gennamen sind durch eggNOG mapper zugewiesen.

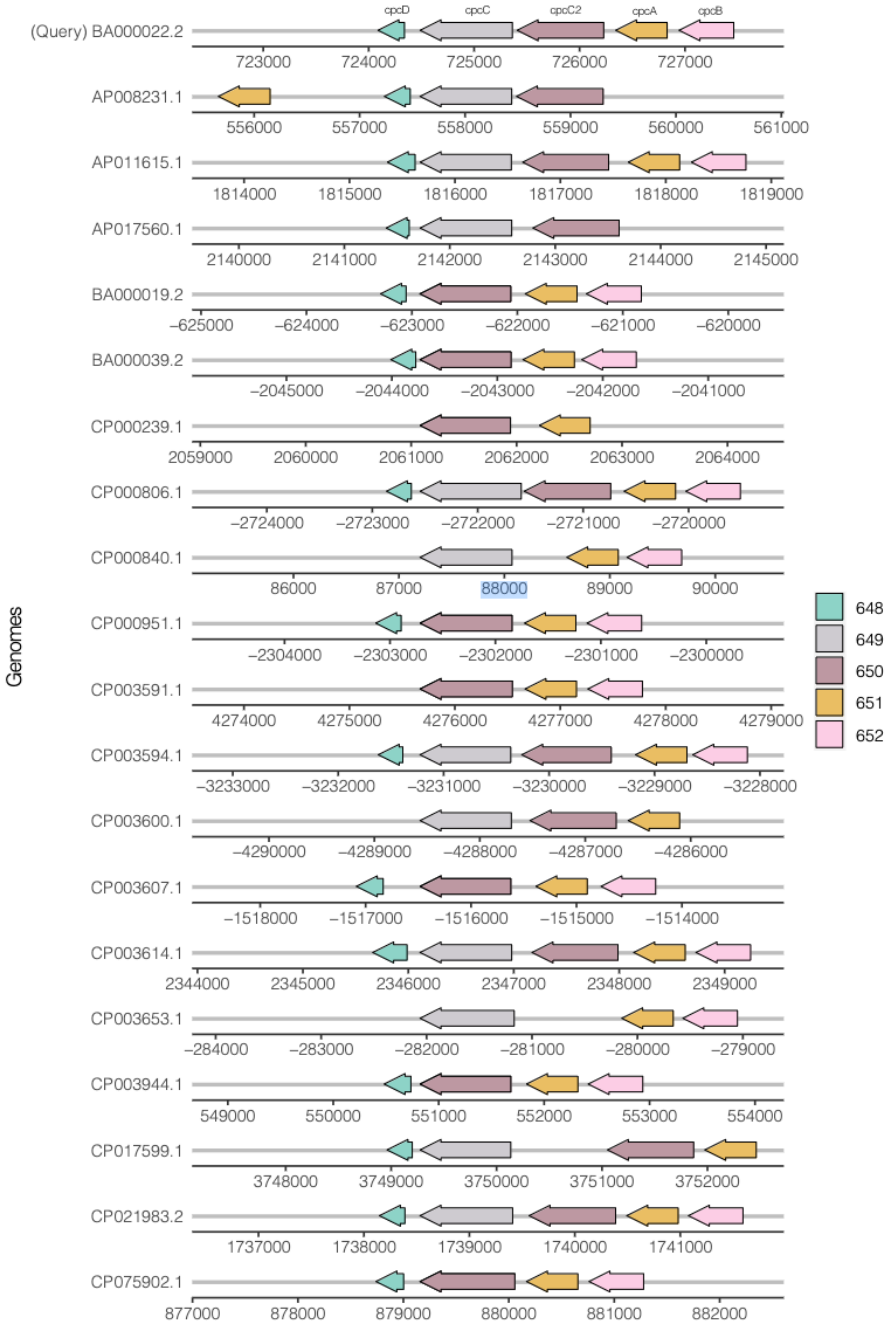

Abb. 10: Durch Spacedust identifiziertes Gencluster für den Phycobilosome Rod Complex. Orthologe Gene/Proteine sind gleich gefärbt. Gennamen sind durch eggNOG mapper (Cantalpiedra 2021) zugewiesen.

2.7. Zurverfügungstellung eines gut dokumentierten open-source-Software-Pakets für Spacedust sowie eines Datensatzes für funktionelle Annotationen der analysierten Referenzgenome

Spacedust ist in C++ implementiert und steht als Open-Source quelloffene benutzerfreundliche Kommandozeilen-Software mit Gnu Public License für Linux und macOS zur Verfügung. Der Spacedust-Quellcode und ein Benutzerhandbuch sind verfügbar unter https://github.com/soedinglab/Spacedust.

Um Spacedust möglichst leicht zugänglich zu machen, haben wir eine Google Colaboratory-Umgebung eingerichtet, die es unseren Nutzern ermöglicht, Tests durchzuführen und Ergebnisse zu reproduzieren, ohne dass eine lokale Installation oder Konfiguration erforderlich ist. Wir stellen ein umfassendes IPython-Notebook (ipynb-Datei) zur Verfügung, das die Schritte zur Installation aller Abhängigkeiten und Datenbanken, das Ausführen des Programms und die interaktive Visualisierung von Clustern umfasst. Innerhalb des Colab Framework können Benutzer Spacedust entweder in einem Alle-gegen-Alle-Modus ausführen, oder sie können Genome mit Hilfe der vorkompilierten Referenzdatenbank mit nur einem einzigen Klick annotieren.

Spacedust ist bereits in vielen Labs rund um die Welt in Benutzung, wie die 42 Sterne für Spacedust zeigen (https://github.com/soedinglab/spacedust/stargazers).

Der Datensatz der 1308 Referenzgenome mit den von Spacedust identifizierten 107 Millionen Genclustermatches ist verfügbar unter dem Link https://wwwuser.gwdguser.de/~compbiol/spacedust/.

3. Wichtigste Positionen des zahlenmäßigen Nachweises

Die Kostenplanung wurde eingehalten. Folgende Kosten sind angefallen:

Kostenart	**Summe der Ausgaben:**
Beschäftigte E12-E15	450.113,69 €
Lohnempfänger(innen) / Sonstige	14.400,00 €
Sonstige allgemeine Verwaltungsausgaben	46.451,37 €
Dienstreisen	- €
High-Memory-Server	29.635,94 €
Entstandene Ausgaben insgesamt	**540.601,00 €**

4. Notwendigkeit und Angemessenheit der geleisteten Arbeit

Die im Forschungsprojekt horizontal4meta durchgeführten Forschungsarbeiten und die dafür aufgewandten Ressourcen entsprechen der im Projektantrag formulierten

Planung und waren, wie dort erläutert, für die Erreichung der gesteckten Ziele notwendig. Eine streng nach ökonomischen Gesichtspunkten durchgeführte Kosten-Nutzen-Analyse ist zu diesem Zeitpunkt nicht möglich. Zwar lässt sich der Nutzen einer wissenschaftlichen Software gut and Zitationen und Downloads pro Zeit messen, jedoch erreichen diese Leistungsmasse besonders bei sehr erfolgreicher Software ihre Maximalwerte erst viele Jahre nach der Veröffentlichung der Software. Als Beispiel sollen hier unsere Software HHpred, HHblits, MMseqs2 und Foldseek dienen (Abbildung 11).

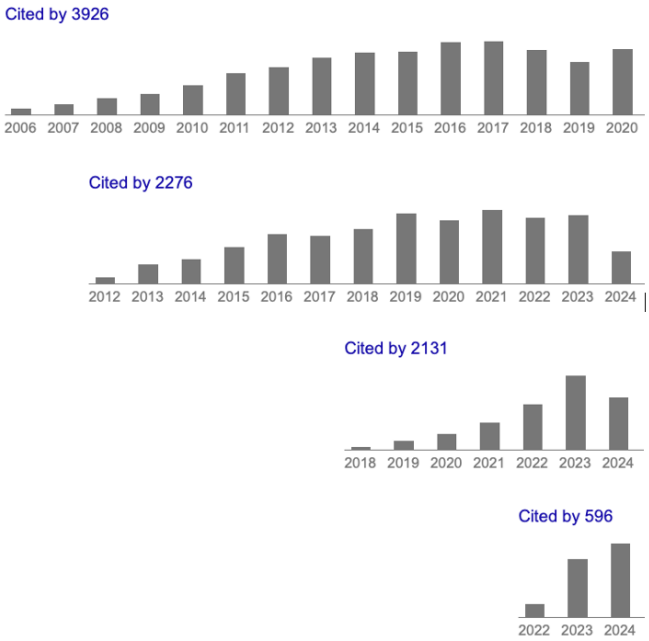

Abb. 11: Anzahl der Zitationen von HHpred (Söding 2006), HHblits (Remmert 2012), MMseqs2 (Steinegger 2017), und FoldSeek (van Kempen 2023). Der Maximalwert der Zitationen pro Jahr wird bei erfolgreicher Software erst viele Jahre nach deren Einführung erreicht.

Es ist davon auszugehen, dass Spacedust und weitere Software und Datenbanken (z.B. PFMOD, siehe Teil III), die auf dem horizontal4meta-Projekt aufbauen werden, ähnlich wie unsere hier als Beispiele aufgeführten Softwarepakete, ebenfalls tausende Zitate bekommen werden. Unsere Erwartung ist, dass Spacedust und darauf aufbauende Software und Datenbanken einen kritischen Beitrag leisten werden zum besseren Verständnis der Mikrobiome im menschlichen Darm und in dutzenden weiteren, für die medizinische, biotechnologische und ökologische Forschung hochrelevanten Lebensräumen.

5. Verwertbarkeit der Ergebnisse

Die im horizontal4meta-Vorhaben entwickelte Software mmseqs taxonomy ist bereits 151-mal zitiert, die SpacePharer-Software bereits 57-mal. Auch wird unsere neueste in horizontal4meta entwickelte Software Spacedust schon von vielen internationalen

Nutzern verwendet, siehe https://github.com/soedinglab/spacedust/stargazers. Die quelloffene, freie UniOP-Software ist seit kurzem ebenfalls öffentlich zugänglich (https://github.com/hongsua/UniOP) und hat bereits erste Nutzer gewinnen können (https://github.com/hongsua/UniOP/stargazers). Die Ergebnisse des horizontal4meta-Projekts werden also bereits verwertet. Allerdings wird, wie oben beschrieben, der Gesamtnutzen erst nach einiger Zeit einschätzbar sein.

6. Fortschritte auf dem Gebiet des Vorhabens bei anderen Stellen

Die Dringlichkeit für die Entwicklung besserer Software zur Annotation mikrobieller Genome hat sich durch die Beschleunigung des Wachstums metagenomischer Datenbanken und der Anzahl metagenomisch assemblierter Genome weiter verschärft. Einen wesentlichen Einfluss auf die funktionale Annotation mikrobieller Genome hat sich ergeben aus der allgemeinen Verfügbarkeit von Strukturen für praktisch jedes Protein, die durch den Durchbruch bei der Proteinstrukturvorhersage mit AlphaFold2 möglich wurde. Die Sensitivität der Homologieinferenz wird sich durch Ersetzen von Sequenz-basierter mit Proteinstruktur-basierter Suche verbessern. (Hieran hat unsere Software FoldSeek einen wichtigen Anteil.) Diese Entwicklung nutzt Spacedust aber bereits aus, weil es bereits FoldSeek-Suchen verwendet. Der Spacedust-Ansatz der horizontalen Annotationstransfers ist außerdem völlig komplementär zum vertikalen Annotationstransfer durch Homologieinferenz.

7. Erfolgte und geplante Veröffentlichungen

Mirdita M, Steinegger M, Breitwieser F, Söding J, Levy Karin E. Fast and sensitive taxonomic assignment to metagenomic contigs. Bioinformatics. 2021; 37(18):3029-31. Zitiert: 135-mal.

Zhang R, Mirdita M, Levy Karin E, Norroy C, Galiez C, Söding J. SpacePHARER: sensitive identification of phages from CRISPR spacers in prokaryotic hosts. Bioinformatics. 2021; 37(19):3364-6. Zitiert: 57-mal.

Su H, Zhang R, Söding J. UniOP: A universal operon prediction for high-throughput prokaryotic (meta-)genomic data. 2024; In preparation.

Zhang R, Mirdita M, Söding J. De novo discovery of conserved gene clusters in microbial genomes with Spacedust. 2024. In preparation.

Kolodyazhnaya A, Oruc YB, Zhang R, Söding J. Large-scale, self-supervised discovery and annotation of functional modules in microbial genomes. 2024. In preparation.

Literaturverzeichnis

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Molecular Biology and Evolution. 2021 Dec 1;38(12):5825-9.

Jahangiri-Tazehkand S, Wong L, Eslahchi C. OrthoGNC: a software for accurate identification of orthologs based on gene neighborhood conservation. Genomics, Proteomics and Bioinformatics. 2017 Dec;15(6):361-70.

Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Research. 2023 Jan 6;51(D1):D587-92.

Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods. 2012 Feb;9(2):173-175.

Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology. 2017 Nov;35(11):1026-8.

Taboada B, Estrada K, Ciria R, Merino E. Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes. Bioinformatics. 2018 Dec 1;34(23):4118-20.

van Kempen M, Kim SS, Tumescheit C, Mirdita M, Gilchrist CL, Söding J, Steinegger M. Foldseek: fast and accurate protein structure search. Nature Biotechnology. 2022.

Westover BP, Buhler JD, Sonnenburg JL, Gordon JI. Operon prediction without a training set. Bioinformatics. 2005 Apr 1;21(7):880-8.

Winter S, Jahn K, Wehner S, Kuchenbecker L, Marz M, Stoye J, Böcker S. Finding approximate gene clusters with Gecko 3. Nucleic Acids Research. 2016 Nov 16;44(20):9600-10.