



ZENTRUM FÜR BIOINFORMATIK (ZBH)
UNIVERSITÄT HAMBURG
HAMBURG, DEUTSCHLAND

PROJEKT GENOMINFORMATIK

Ein systematischer Vergleich von
Verfahren zur funktionellen und
taxonomischen Klassifikation von
metagenomischen
Sequenzfragmenten

Marie Sofie Briem, Inga Lemme, Sarah Weber

Gutachter/in
Prof. Dr. Kurtz, Dr. Gonella

23. Februar 2016

Inhaltsverzeichnis

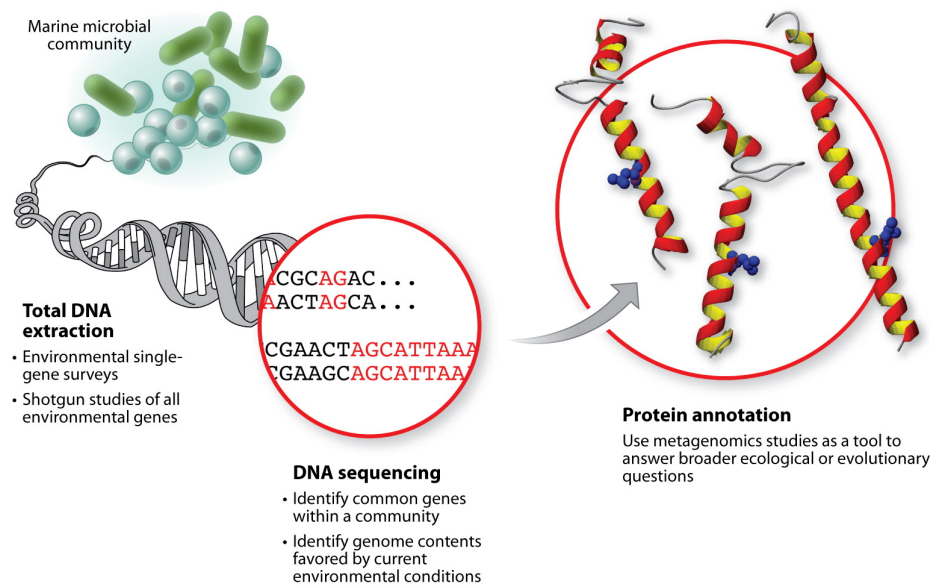
1	Einleitung	4
1.1	Diamond	5
1.2	Lambda	5
1.3	Ziele	6
2	Material und Methoden	7
2.1	Material	7
2.1.1	Datensets	7
2.1.2	Datenbank	8
2.1.3	Programme	8
2.2	Methoden	10
2.2.1	Vorarbeit	10
2.2.2	Programmausgabe	10
2.2.3	Megan	10
2.2.4	Newick	10
2.2.5	Genauigkeitsberechnungen	10
3	Ergebnisse	11
3.1	Distanzverteilung	11
3.2	Sensitivität und Präzision	17
3.3	Laufzeitverhalten	19
4	Diskussion	20
4.1	Distanzverteilung, Sensitivität und Präzision	20
5	Literaturverzeichnis	21

Abbildungsverzeichnis

1	Metagenomik in Schritten	4
2	Spaced Seeds	5
3	Distanzverteilung der Reads: Carma Datensatz.	11
4	Distanzverteilung der Reads: FACS Datensatz.	12
5	Distanzverteilung der Reads: PhyloPythia Datensatz.	13
6	Distanzverteilung der Reads: Metaphyler Datensatz.	14
7	Distanzverteilung der Reads: PhymmBL Datensatz.	15
8	Distanzverteilung der Reads: RAlphy Datensatz.	16

1 Einleitung

Der Forschungsbereich der Metagenomik beschäftigt sich mit der Klassifizierung und Zuordnung aller genetischen Informationen, die in zufällig entnommenen Proben enthalten sind [7]. Die Proben bestehen beispielsweise aus marinen mikrobiellen Wasser- oder Bodenproben, mit Hilfe derer ökologische Fragestellungen beantwortet werden sollen. Ein weiteres bedeutsames Forschungsgebiet der Metagenomik ist die Beschäftigung mit dem humanen Mikrogenom welches wichtige Informationen unter anderem zur Ernährung, Regulation des Immunsystem und der Aufklärung von Krankheitsresistenzen geben kann [14]. Aus diesen Proben wird die gesamte enthaltene DNA extrahiert und sequenziert. Anschließend werden die Proteine annotiert um Funktion und taxonomische Zuordnung der enthaltenden Spezies zu ermitteln (Abb. 1). Basierend auf neuen und schnellen Sequenzierungstechnologien wie Illumina, fallen im Bereich der Metagenomik große Datenmengen an, die taxonomisch und funktionell klassifiziert werden müssen. Eine vielversprechende Alternative zu dem Alignierprogramm BlastX, welches mithilfe von Sequenzvergleichen eine solche Klassifizierung durchführt, scheinen Diamond [5] und Lambda [8] zu sein, die eine Laufzeitersparnis mit Hilfe von double indexing erwirken sollen.




 Gilbert JA, Dupont CL. 2011.
Annu. Rev. Mar. Sci. 3:347–71

Abbildung 1: Metagenomik in Schritten

1.1 Diamond

Das open source verfügbare Alignierprogramm Diamond [5] basiert auf einem Seed- und Extent Algorithmus. Im Seedingschritt werden so gennante Spaced Seeds gesucht, die als Treffer in Anfrage- und Datenbanksequenz gefunden werden sollen (Abb. 2). Das Seeding findet anschließend mit Double Indexing statt. Beim Double Indexing werden sowohl Anfrage- als auch Referenzdatenbanksequenz geindext, was eine geringere Laufzeit durch schnelleres durchsuchen der Datenstrukturen mit sich bringt. Die Indizierung findet bei Diamond "on the fly", das heißt während des Programmdurchlaufs, statt. Die Treffer, die mit Hilfe der Spaced Seeds gefunden wurden, speichert Diamond in lexikographisch sortierten Listen.

```
(a) 111101011101111
     111011001100101111
     1111001001010001001111
     111100101000010010010111

(b) Reference    SLWAKKRTVDGQPKWLPLVAHLVDASNVSRMLFNQWLS
     Spaced seed  111101011101111
     Query       FWAKKRTNDGQKWLPLTQHLEDASNVSR
```

Abbildung 2: Spaced Seeds

Um eine Erweiterung im Extendschritt durchzuführen, überprüft das Programm, ob der Seed-Treffer größer-gleich 10 Aminosäuren lang ist. Der Seed wird schließlich mit dem Smith-Waterman Algorithmus erweitert.

1.2 Lambda

Auch Lambda [8] basiert wie Diamond auf dem Seed- und Extend Algorithmus mit Double Indexing. Im Gegensatz zu Lambda muss die Referenzdatenbank vorgeindext werden und findet nicht während des Programmdurchlaufs statt. Als Datenstrukturen stehen für die Referenzdatenbank ein Suffixarray und für die Anfragesequenz ein Radixtree zur Verfügung. Die Speicherung in einem Radix-tree ermöglicht eine Paralellisierung verschiedener Seeds, was eine zusätzliche Zeitersparnis bedeutet. Die Erweiterung erfolgt mit Hilfe des X-drop Algorithmus.

1.3 Ziele

Das Projekt hat folgende Ziele:

- Bestimmung und Vergleich der Sensivität (Anzahl an korrekt Bestimmten / Anzahl an Sequenzen im Datenset) und Präzision (Anzahl an korrekt Bestimmten / Anzahl an Zugewiesenen) der Programme Diamond und Lambda.
- Bestimmung und Vergleich der benötigten Zeit der Programme Diamond und Lambda.

2 Material und Methoden

2.1 Material

Das durchgeführte Projekt orientiert sich an der Forschungsarbeit von Bazinet und Cummings [3]. Im Folgenden bezieht sich der Ausdruck "Vorlagepaper" auf die Arbeit von Bazinet und Cumming. Die Datensets und Datenbanken wurden anhand des Vorlagepapers ausgewählt.

2.1.1 Datensets

Die Experimente wurden mit folgenden Datensets durchgeführt:

1. *FACS 269bp* – Datenset, original von Strannenheim *et al.* [13], bestehend aus **27.049 simulierten 454 Reads** mit einer durchschnittlichen Länge von **269 bp**. Die im Vorlagepaper angegebene Referenz war nicht mehr aktuell und konnte nicht gefunden werden. Das Datenset wurde direkt von Herrn Bazinet bereit gestellt. Das bereitgestellte Datenset beinhaltet 72.951 Reads der Spezies *Homo sapiens*. Diese Reads wurden entfernt, so dass das genutzte Datenset 27.049 Reads enthält. Das Datenset setzt sich aus 19 bakteriellen und drei viralen Genomen zusammen [3].
2. *CARMA 265bp* – Datenset bestehend aus **25.000 simulierten 454 Reads** mit einer durchschnittlichen Länge von **265 bp**, original genutzt von Gerlach und Stoye [6]. Das Datenset wurde von der WebCARMA Homepage unter dem Link http://www.cebitec.uni-bielefeld.de/webcarma.cebitec.uni-bielefeld.de/download/simulated_metagenome_454_265bp.fna heruntergeladen. Zusammengesetzt ist das Datenset aus 25 bakteriellen Genomen, die sich wie folgt in die einzelnen bakteriellen Phyla verteilen: 73,0% Proteobacteria; 12,9% Firmicutes; 7,8% Cyanobacteria; 5,2% Actinobacteria; 1,0% Chlamydiae [3].
3. *Metaphyler 300bp* – Datenset bestehend aus **73.086 simulierten Reads** von 31 phylogenetischen Markern bakterieller Genome mit einer durchschnittlichen Länge von **300 bp**. Ursprünglich genutzt von Liu *et al.* [10]. Das Datenset konnte anhand der im Vorlagepaper angegebenen Referenz nicht korrekt ermittelt werden. Die Rechercheergebnisse ergaben ein Datenset bestehend aus 40.039 Reads mit einer durchschnittlichen Länge von 645 bp. Das korrekte Datenset wurde direkt von Herrn Bazinet bereit gestellt. Die Verteilung in die bakteriellen Phyla setzt sich folgendermaßen zusammen: 47.0% Proteobacteria; 21.9% Firmicutes; 9.7% Actinobacteria; 4.8% Bacteroidetes; 3.9% Cyanobacteria; 2.2% Tenericutes; 1.9% Spirochaetes; 1.3% Chlamydiae; 0.9% Thermotogae; 0.9% Chlorobi [3].

4. *PhymmBL 243bp* – Datenset bestehend aus **80.215 RefSeq Reads** mit einer durchschnittlichen Länge von **243 bp**. Das Datenset, original genutzt von Brady und Salzberg [4], konnte anhand der im Vorlagepaper angegebenen Referenz nicht korrekt ermittelt werden. Die Rechercheergebnisse ergaben ein Datenset bestehend aus 73.252 Reads mit einer durchschnittlichen Länge von 204 bp. Das korrekte Datenset wurde direkt von Herrn Bazinet bereit gestellt.
5. *PhyloPythia 969bp* – simMC Datenset bestehend aus **114.457 Reads** mit einer durchschnittlichen Länge von **969 bp**, original von Patil *et al.* [12]. Das im Vorlagepaper verwendete Datenset "PhyloPythia" bestehend aus 124.941 Reads mit einer durchschnittlichen Länge von 961 bp konnte auch mit Hilfe von Herrn Bazinet nicht ermittelt werden. Das Datenset wurde von der JGI Homepage unter dem Link http://fames.jgi-psf.org/Retrieve_data.html heruntergeladen.
6. *RAIphy 233bp* – Datenset bestehend aus **477.000 RefSeq Reads** mit einer durchschnittlichen Länge von **233 bp**, original von Nalbantoglu *et al.* [11]. Das im Vorlagepaper verwendete Datenset "RAIphy" bestehend aus 477.000 Reads mit einer durchschnittlichen Länge von 238 bp konnte mit der angegebenen Referenz nicht ermittelt werden. Das im Projekt verwendete Datenset wurde von Herrn Bazinet zur Verfügung gestellt.

2.1.2 Datenbank

Für die Suche wurde die Datenbank UniProtKB/Swiss-Prot von UniProt verwendet [2]. Diese wurde unter dem Link <http://www.uniprot.org/downloads> heruntergeladen. Die Datenbank besteht aus 549.646 Sequenzen mit einer durchschnittlichen Länge von 356.56bp.

2.1.3 Programme

Im Projekt wurden folgende Programme verwendet:

1. **Lambda** – Das Programm Lambda (Version 0.9.2) wurde von der GitHub Seite mit dem Link <https://github.com/seqan/lambda.git> heruntergeladen [8]. Dabei wurde wie folgt vorgegangen:

```
$ git clone https://github.com/seqan/lambda.git
$ cd lambda
$ mkdir build
```



```
$ cmake -DCMAKE_C_COMPILER=/usr/local/zbhtools/gcc/gcc-5.1.0/bin/gcc
-DCMAKE_CXX_COMPILER=/usr/local/zbhtools/gcc/gcc-5.1.0/bin/g++
-DCMAKE_INSTALL_PREFIX=/work/gi/software
$ make -j2
```

Um das Programm Lambda ausführen zu können musste die UniProtKB/Swiss-Prot Datenbank zunächst indiziert werden. Dazu wurde das Programm "lambda_indexer" verwendet, welches in dem oben genannten Packet enthalten ist. Folgender Aufruf wurde verwendet:

```
$ lambda_indexer -d uniprot_sprot.fasta
```

Die jeweiligen Datensets (s.o.) wurden gegen die indizierte Datenbank mit dem Befehl

```
$ lambda -q QUERY.fasta -d DATABASE.fasta [-o output.m8]
```

aligniert.

2. **Diamond** – Das Programm Diamond (Version 0.7.9) wurde von der GitHub Seite mit dem Link <https://github.com/bbuchfink/diamond>. git heruntergeladen [5]. Es wurde folgendermaßen verfahren:

```
$ git clone https://github.com/bbuchfink/diamond.git
$ cd diamond
$ mkdir build
$ cmake -DCMAKE_INSTALL_PREFIX=/work/gi/software/diamond
$ make install
```

Mit dem Aufruf

```
$ diamond makedb -in uniprot_sprot.fasta -d diamonduniprot_sprot.fasta.dmnd
```

wurde die binäre Diamond-Datenbank aus der UniProtKB/Swiss-Prot Datenbank erstellt.

Die jeweiligen Datensets (s.o.) wurden gegen die zuvor erstellte Diamond-Datenbank mit dem Befehl

```
$ diamond blastx -d DIAMOND_DATABASE.dmnd -q QUERY.fasta
-a OUTPUT -t <temporary directory>
```

aligniert.

3. **MEGAN5** – Das Programm MEGAN5 (Version 5.10.7) wurde von der Website <http://ab.inf.uni-tuebingen.de/data/software/megan5/download/welcome.html> heruntergeladen [9]. Die benötigte akademische Lizenz wurde uns von Herrn Dr. Giorgio Gonella zur Verfügung gestellt. MEGAN5 ist ein Programm mit einer graphischen Benutzeroberfläche. Das Programm benötigt eine "map" gegen die es die Eingabereads vergleicht. Es konnte nicht die voreingestellte "map" genutzt werden, da diese gegen GI-Nummern sucht, die Ausgabe von Lambda und Diamond jedoch durch die Nutzung der UniProtKB/Swiss-Prot Datenbank keine GI-Nummern sondern sp-Nummern generiert. Aus diesem Grund wurde die "idmapping.dat" von der Website ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/ heruntergeladen. Diese wurde anschließend gekürzt, so dass nur die Taxa enthalten waren, die auch in der UniProtKB/Swiss-Prot Datenbank vorkommen.

2.2 Methoden

2.2.1 Vorarbeit

2.2.2 Programmausgabe

2.2.3 Megan

2.2.4 Newick

2.2.5 Genauigkeitsberechnungen

3 Ergebnisse

3.1 Distanzverteilung

Abbildungen 3 - 8 zeigen die Distanzverteilungen der Reads, welche mithilfe des newick-parser ermittelt wurden. Um einen direkten Vergleich von Diamond und Lambda vornehmen zu können, wurden pro Datensatz die Ergebnisse beider Programme gegeneinander gestellt. Die Verteilung der Datensätze Carma (Abb. 3), FACS (Abb. 4) und PhyloPythia (Abb. 5) zeigt die gemeinsame Tendenz, dass die Ausgabe des Programms Diamond eine hohe Anzahl Reads mit einer geringen Distanz aufweist. Die Ausgaben von Lambda zeigen erst bei einer Distanzgröße von größer als 4 bemerkenswerte Readanzahlen.

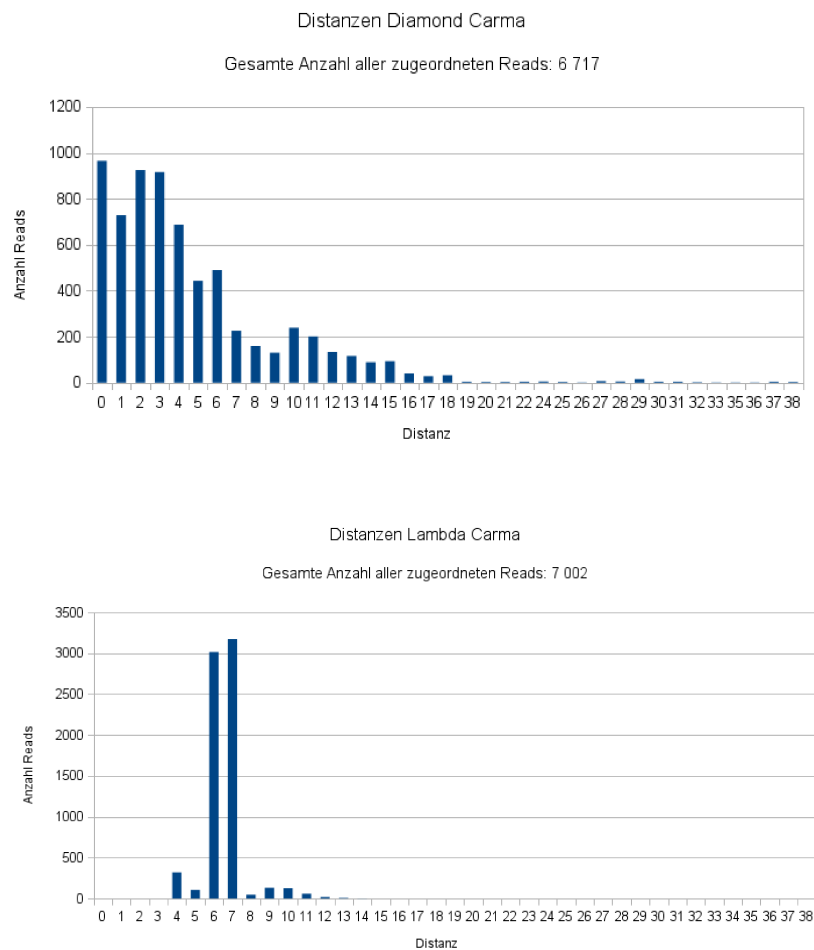


Abbildung 3: Distanzverteilung der Reads: Carma Datensatz.
Oben: Ausgabe Diamond. **Unten:** Ausgabe Lambda.

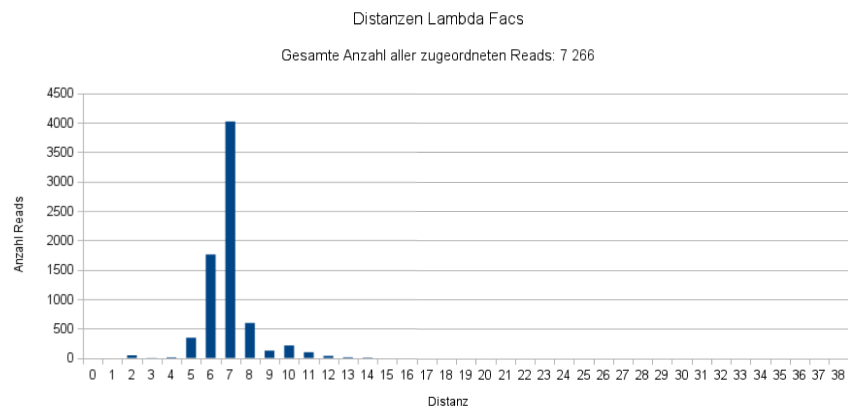
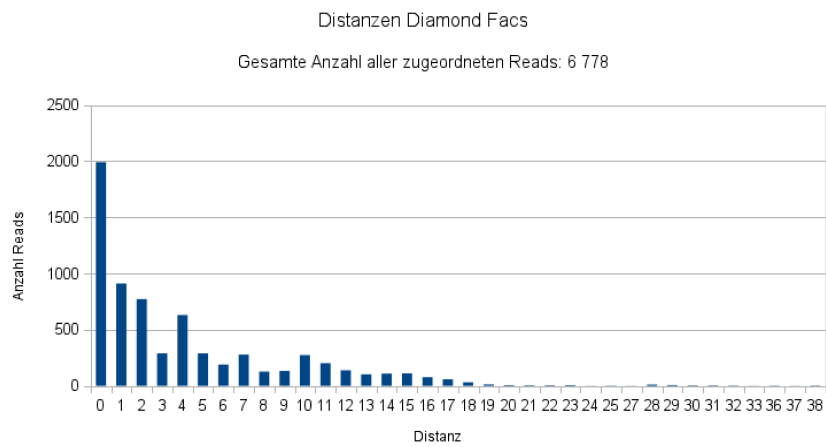


Abbildung 4: Distanzverteilung der Reads: FACS Datensatz.
Oben: Ausgabe Diamond. **Unten:** Ausgabe Lambda.

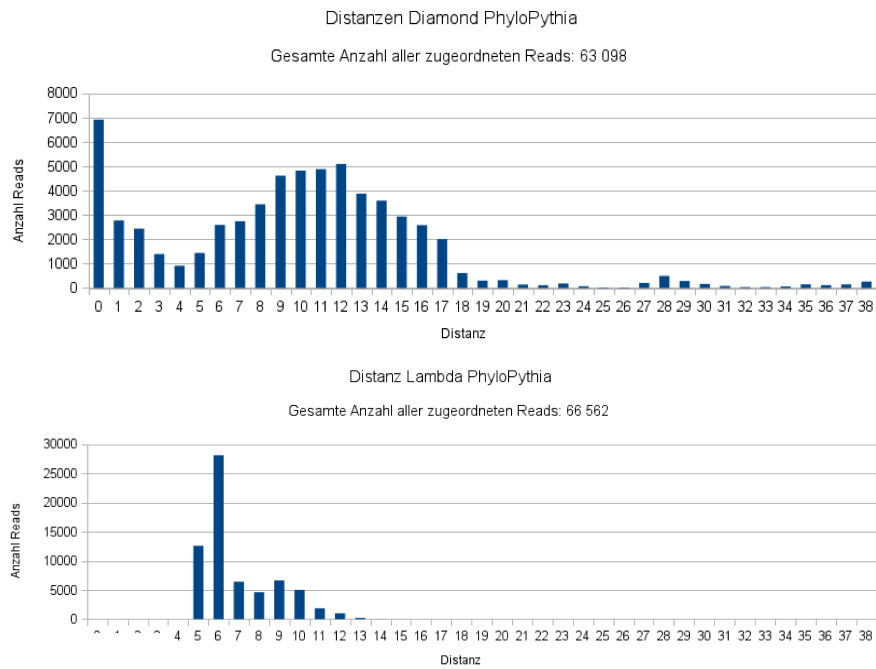


Abbildung 5: Distanzverteilung der Reads: PhyloPythia Datensatz.
Oben: Ausgabe Diamond. **Unten:** Ausgabe Lambda.

Die jeweiligen Ergebnisse der beiden Programme für die übrigen drei Datensätze (Abb. 6-8) ähneln sich sehr. Auffällig ist hier, dass die Zuweisung der Reads bei beiden Programmen für die Datensätze Metaphyler (Abb. 6) und PhymmBL (Abb. 7) erst nennenswerte Readanzahlen bei einer Distanz von größer als 4 erzeugen. Für den Datensatz RAIphy können bei beiden Programmen hohe Readanzahlen bei einer Distanz von 0 beobachtet werden.

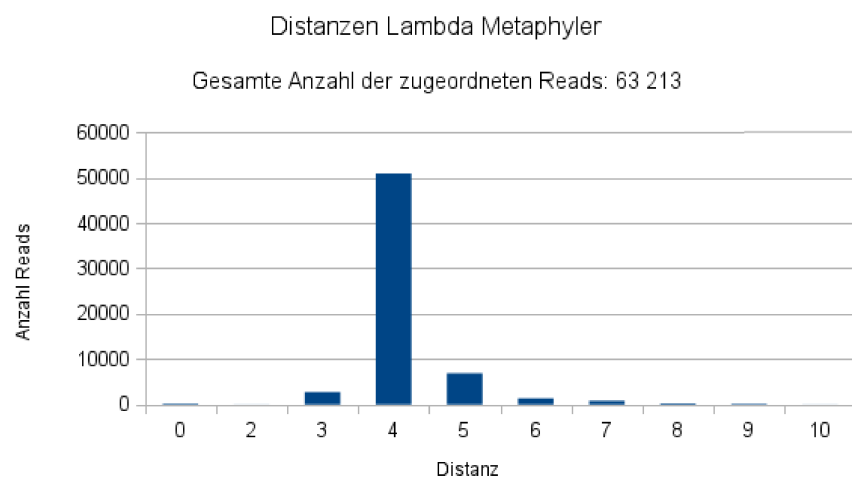
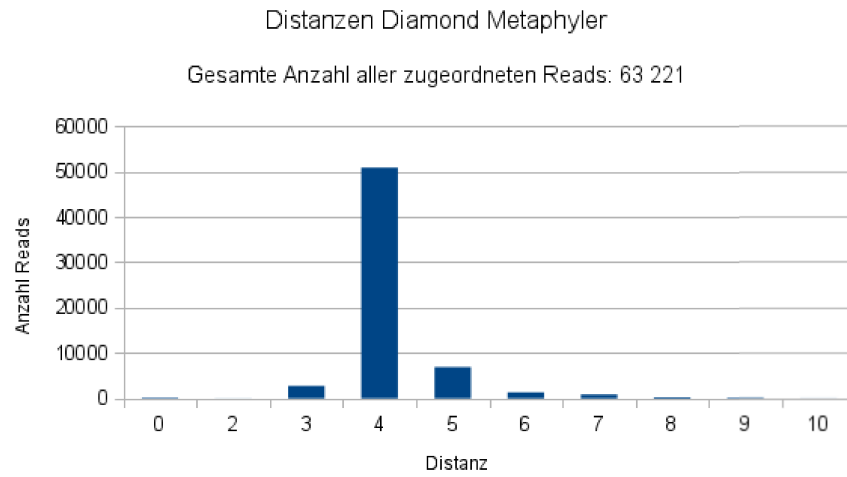


Abbildung 6: Distanzverteilung der Reads: Metaphyler Datensatz.
Oben: Ausgabe Diamond. **Unten:** Ausgabe Lambda.

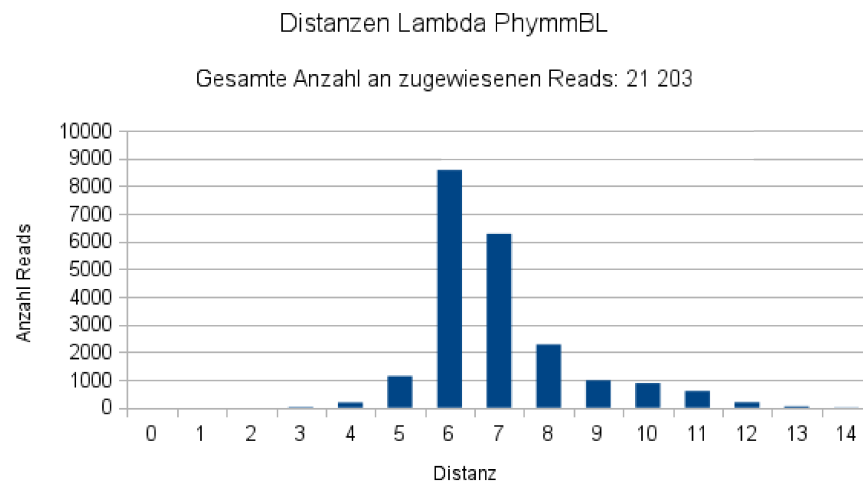
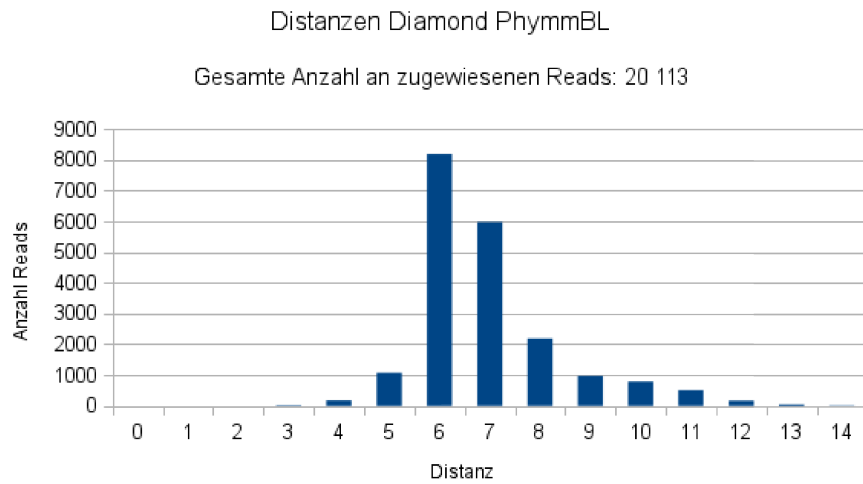


Abbildung 7: Distanzverteilung der Reads: PhymmBL Datensatz.
Oben: Ausgabe Diamond. **Unten:** Ausgabe Lambda.

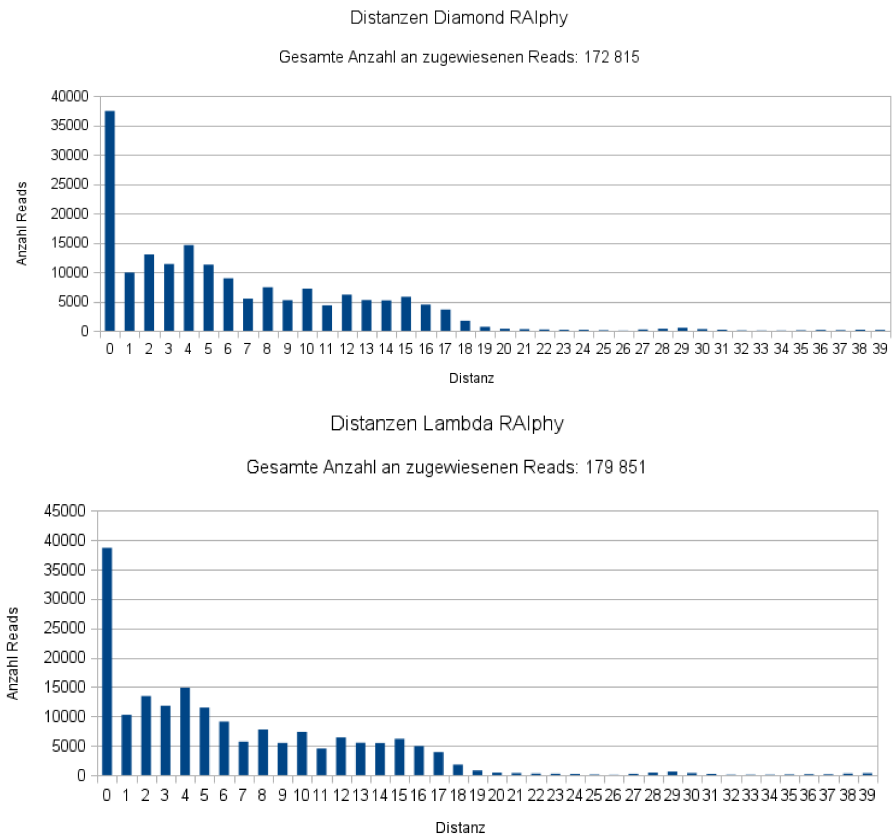


Abbildung 8: Distanzverteilung der Reads: RAIphy Datensatz.
Oben: Ausgabe Diamond. **Unten:** Ausgabe Lambda.

Für alle Datensätze lässt sich erkennen, dass Lambda insgesamt mehr Reads zugeordnet hat als Diamond.

3.2 Sensitivität und Präzision

Die oben beschriebene Tendenz (Kapitel 3.1) zeigt sich auch für die Genauigkeitsberechnungen. Die ausgewertete Lambda-Ausgabe zeigt für den Datensatz FACS (Tab. 2) erst bei Distanzen ab 2 nennenswerte Sensitivität- und Präzisionswerte, Carma (Tab. 1) und PhyloPythia (Tab. 3) sogar erst ab Distanzen von 4 - 5. Diamond dagegen weist bei den drei genannten Datensätzen schon bei einer Distanz von 0 auswertbare Genauigkeitsergebnisse auf.

Distanz	0	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5
Programm						
Lambda	0	0	0	0	0,0128	0,0170
Diamond	0,0386	0,0677	0,1047	0,1413	0,1688	0,1865

Distanz	0	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5
Programm						
Lambda	0	0	0	0	0,0456	0,0606
Diamond	0,1437	0,2521	0,3898	0,5262	0,6284	0,6944

Tabelle 1: Genauigkeitsberechnung der Reads: Carma Datensatz.
Oben: Sensitivität. **Unten:** Präzision.

Distanz	0	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5
Programm						
Lambda	0	0	0,0016	0,0017	0,0021	0,0148
Diamond	0,0736	0,0107	0,1358	0,1464	0,1696	0,1802

Distanz	0	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5
Programm						
Lambda	0	0	0,0060	0,0063	0,0078	0,0550
Diamond	0,2936	0,4280	0,5418	0,5842	0,6770	0,7194

Tabelle 2: Genauigkeitsberechnung der Reads: FACS Datensatz.
Oben: Sensitivität. **Unten:** Präzision.

Distanz	0	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5
Programm						
Lambda	0	0	0	0	0	0,1101
Diamond	0,0605	0,0847	0,1060	0,1182	0,1262	0,1388

Distanz	0	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5
Programm						
Lambda	0	0	0	0	0	0,1892
Diamond	0,1098	0,1538	0,1925	0,2146	0,2291	0,2519

Tabelle 3: Genauigkeitsberechnung der Reads: PhyloPythia Datensatz.
Oben: Sensitivität. **Unten:** Präzision.

Die drei übrigen Datensätze Metaphyler (Tab. 4), PhymmBL (Tab. 5) und RAIphy (Tab. 6) zeigen sowohl bei der Sensitivität, als auch bei der Präzisionsberechnung für Diamond und Lambda nahezu identische Ergebnisse.

Distanz	0	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5
Programm						
Lambda	0,0004	0,0004	0,0004	0,0074	0,1380	0,1557
Diamond	0,0004	0,0004	0,0004	0,0074	0,1378	0,1555

Distanz	0	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5
Programm						
Lambda	0,00270	0,0270	0,0029	0,0459	0,8504	0,9594
Diamond	0,00260	0,0026	0,0028	0,0462	0,8498	0,9588

Tabelle 4: Genauigkeitsberechnung der Reads: Metaphyler Datensatz.
Oben: Sensitivität. **Unten:** Präzision.

Distanz	0	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5
Programm						
Lambda	0	0	0	0,0003	0,0026	0,0168
Diamond	0	0	0	0,0003	0,0025	0,0159

Distanz	0	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5
Programm						
Lambda	0	0	0	0,0011	0,0100	0,0637
Diamond	0	0	0	0,0011	0,0100	0,0635

Tabelle 5: Genauigkeitsberechnung der Reads: PhymmBL Datensatz.
Oben: Sensitivität. **Unten:** Präzision.

Distanz Programm	0	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5
Lambda	0,0811	0,1026	0,1308	0,1555	0,1867	0,2108
Diamond	0,0785	0,0993	0,1266	0,1504	0,1810	0,2046

Distanz Programm	0	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5
Lambda	0,2151	0,2721	0,3470	0,4126	0,4954	0,5593
Diamond	0,2167	0,2741	0,3494	0,4152	0,4996	0,5648

Tabelle 6: Genauigkeitsberechnung der Reads: RAIphy Datensatz.
Oben: Sensitivität. **Unten:** Präzision.

3.3 Laufzeitverhalten

Der direkte Vergleich der Laufzeiten von Lambda und Diamond lässt erkennen, dass Diamond insgesamt bis zu drei Mal schneller läuft als Lambda (Tab. 7). Lediglich für die kleinen Datensätze Carma und FACS ist Lambda schneller.

Datensatz	Lambda [s]	Diamond [s]	Lambda [$\frac{s}{Mbp}$]	Diamond [$\frac{s}{Mbp}$]
Carma	37	68	5	10
FACS	39	72	5	10
PhyloPythia	955	280	9	3
Metaphyler	1 255	788	31	19
PhymmBL	362	176	19	9
RAIphy	1 512	538	14	5

Tabelle 7: Laufzeitverhalten der Programme Diamond und Lambda für die jeweiligen Datensätze in Sekunden und Sekunden pro Megabase

4 Diskussion

Lambda und Diamond sind Programme, die eine Alternative zu dem Alignierprogramm BlastX [1] darstellen sollen [8, 5]. Für eine erste Quantifizierung dieser Aussage wurden im Rahmen dieses Projektes beide Programme miteinander verglichen. Angelehnt war der Vergleich auf der Arbeit von Bazinet und Cummings, 2012 [3]. Die Versuche wurden mit den gleichen Datensets durchgeführt, die auch im "Vorlagepaper" verwendet wurden. Es wird jedoch nicht angegeben, wie die Autoren vorgegangen sind um die Präzision, Sensitivität und das Laufzeitverhalten der von ihnen untersuchten Programme zu berechnen. Aus diesem Grund wurde in diesem Projekt eine eigene Lösung der Berechnung der Genauigkeit und des Laufzeitverhaltens der Programme Lambda und Diamond erstellt. Die Ergebnisse lassen sich demnach nicht direkt mit den Ergebnissen von Bazinet und Cummings vergleichen. Es wurde entschieden, dass die Ergebnisse dieses Projektes unabhängig vom "Vorlagepaper" betrachtet werden.

4.1 Distanzverteilung, Sensitivität und Präzision

Die Ergebnisse bezüglich der Distanzverteilung, Sensitivität und Präzision zeigen, dass Diamond genauer ist als Lambda. Diamond aligniert bei allen Datensets (bis auf PhymmBL und Metaphyler) den Großteil der untersuchten Reads mit den dem Gold-Standard entsprechenden Sequenzen (Distanz von 0) und ist zudem für Datensets Carma, FACS und PhyloPythia deutlich sensibler und präziser als Lambda.

Literatur

- [1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." *J. Mol. Biol.* 215:403-410.
- [2] Bairoch, A., Boeckmann, B., Ferro, S., Gasteiger, E. 2004, Swiss-Prot: juggling between evolution and stability. *Brief Bioinform.* 539–55
- [3] A. L. Bazinet and M. P. Cummings. *A comparative evaluation of sequence classification programs*. *BMC Bioinformatics*, 13: p92, 2012.
- [4] Brady A, Salzberg SL: Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* 2009, 6(9):673-U68. 10.1038/nmeth.1358
- [5] Buchfink B., Xie C., Huson D.H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 2014;12:59-60.
- [6] Gerlach W, Stoye J: Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Res* 2011, 39(14):e91. 10.1093/nar/gkr225
- [7] Jo Handelsman, Michelle R. Rondon, Sean F. Brady, Jon Clardy and Robert M. Goodman. *Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products*. *Chemistry & Biology*, 5(10): 245-249, 1998.
- [8] Lambda: the local aligner for massive biological data; Hannes Hauswedell, Jochen Singer, Knut Reinert; *Bioinformatics* 2014 30 (17): i349-i355; doi: 10.1093/bioinformatics/btu439
- [9] Daniel H Huson, Suparna Mitra, Nico Weber, Hans-Joachim Ruscheweyh, and Stephan C Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, 21:1552–1560, 2011.
- [10] Liu B, Gibbons T, Ghodsi M, Pop M: MetaPhyler: Taxonomic profiling for metagenomic sequences. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. , Hong Kong; 2010:95–100.
- [11] Nalbantoglu OU, Way SF, Hinrichs SH, Sayood K: RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC Bioinf* 2011, 12: 41. 10.1186/1471-2105-12-41

- [12] Patil KR, Haider P, Pope PB, Turnbaugh PJ, Morrison M, Scheffer T, McHardy AC: Taxonomic metagenome sequence assignment with structured output models. *Nat Methods* 2011, 8(3):191–192. 10.1038/nmeth0311-191
- [13] Stranneheim H, Kaller M, Allander T, Andersson B, Arvestad L, Lundeberg J: Classification of DNA sequences using Bloom filters. *Bioinformatics* 2010, 26(13):1595–1600. 10.1093/bioinformatics/btq230
- [14] L. Dethlefsen, S. Huse, M. L. Sogin, and D. A. Relman. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol.*, 6:e280, Nov 2008.