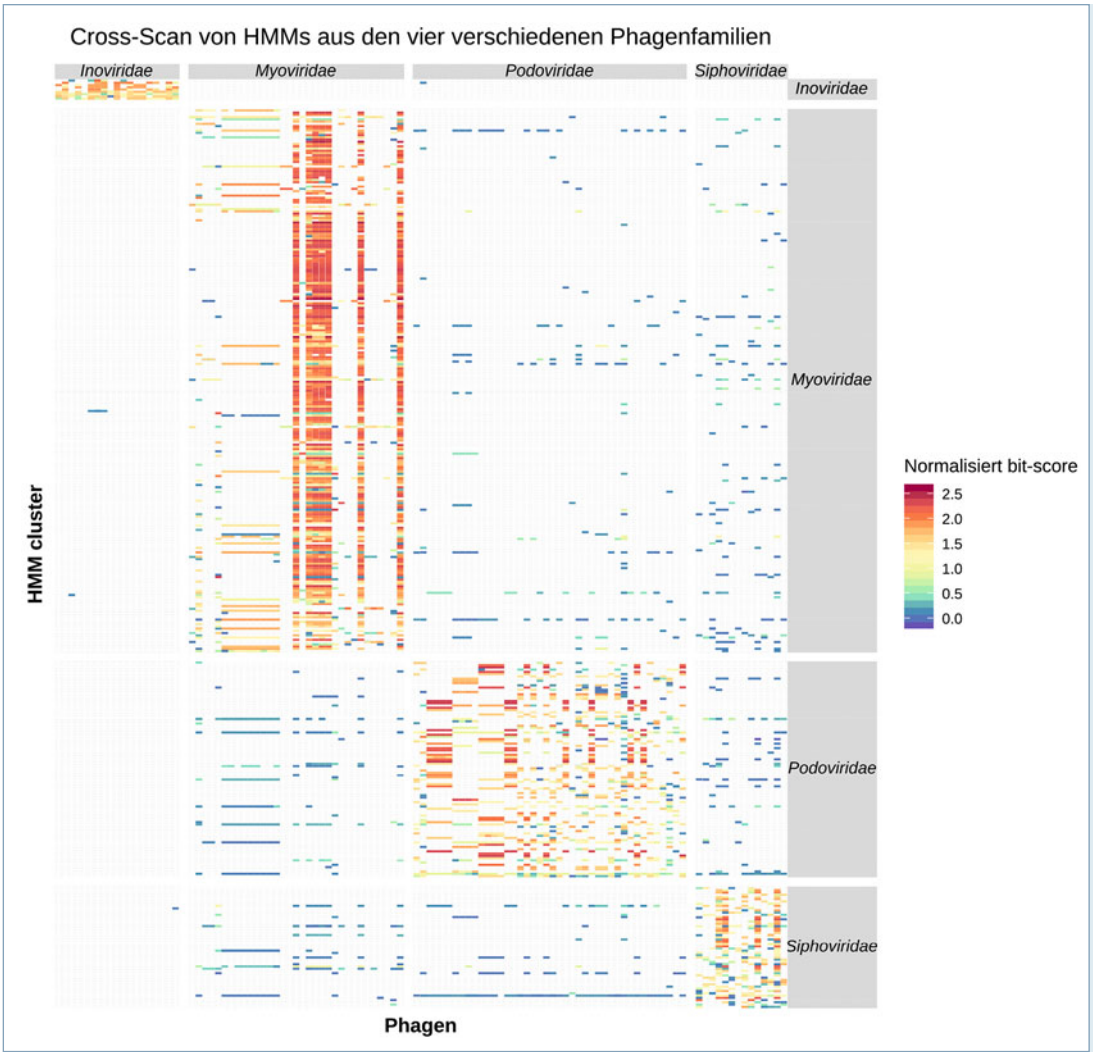
Klassifizierung bakterieller Viren

Phagentaxonomie in der *Next Generation Sequencing*-Ära

CYNTHIA MARIA CHIBANI[1], HEIKO LIESEGANG[2]
[1] INSTITUT FÜR ALLGEMEINE MIKROBIOLOGIE, UNIVERSITÄT ZU KIEL
[2] INSTITUT FÜR MIKROBIOLOGIE UND GENETIK, UNIVERSITÄT GÖTTINGEN

Phages are the biggest known biological entity on earth (about 1031 particles). Due to next generation sequencing methods applied on environmental samples an unpreceeded amount of phage genome data is available. Due to their extreme diversity and the lack of monophyly a sequence based taxonomy is challenging. However, within the phages there are monophyletic subgroups that can be classified based on their genome sequence. A method that combines the shared gene content with taxon specific similarities enables a reliable identification of the phage family based exclusively on the corresponding genome sequence.

DOI: 10.1007/s12268-020-1342-1



◀ **Abb. 1:** Genomscan aller Proteinsequenzen aus Phagengenomen mit taxonspezifischen verdeckten Markov-Modellen (HMM) aus vier Phagenfamilien: Inoviridae, Myoviridae, Podoviridae und Siphoviridae. In den Spalten sind die Treffer farbcodiert dargestellt, mit ansteigender Ähnlichkeit von Blau nach Rot. Um die Ähnlichkeiten unterschiedlich großer Proteine vergleichen zu können, wurden die *bit-scores* der HMM-Treffer über die Länge der Proteine normalisiert. Erkennbar ist, dass sowohl die Zahl als auch die Ähnlichkeit der Treffer eine klare Zuordnung des jeweiligen Phagengenoms zu jeweils einer Phagenfamilie ermöglicht. Die Scans der Myoviridae, Podoviridae und Siphoviridae, die die Gruppe der Caudovirales bilden, zeigen Treffer in den jeweiligen anderen Familien. Diese *cross matches* bestätigen die Verwandtschaft der Genomfamilien, allerdings ist auch hier anhand der eindeutig niedrigeren Ähnlichkeiten dennoch eine Zuordnung zur richtigen Familie möglich (aus [13]).

Tab. 1: Eigenschaften typischer Phagen. Für die Phagenfamilien ist jeweils ein typischer Vertreter beschrieben, die Genomsequenzen sind über den GenBank-Zugriffscode frei verfügbar.

Phage	Familie, genetisches Material, Morphologie	Genomgröße	GenBank
T4	Myovirus, dsDNA, Kopf mit kontraktilem Schwanz	169 kbp, 288 Gene	AF158101.6
T7	Podovirus, dsDNA, Kopf mit kurzem Schwanz	40 kbp, 51 Gene	LR745710.1
I	Siphovirus, dsDNA, Kopf mit nicht-kontraktilem Schwanz	48 kbp, 66 Gene	J02459.1
ΦX174	Microvirus, ssDNA, Kopf	5,4 kbp, 12 Gene	MN385565.1
M13	Inovirus, ssDNA, Filament	6,4 kbp, 10 Gene	V00604.2
Phi6	Cystovirus, dsRNA in drei Segmenten, Kopf	13 kbp, 12 Gene	DQ785293.1, DQ785292.1, DQ785287.1
MS2	Levivirus, ssRNA, Kopf	2,7 kbp, 7 Gene	MH603811.1
PBCV-1	*Giant virus*, dsDNA, Kopf	330 kbp, 802 Gene	JF411744.1

■ Phagen oder auch Bakteriophagen wurden unabhängig voneinander 1915 von Frederick Twort [1] und 1917 von Félix d'Herelle beschrieben [2]. Letzterer prägte den Begriff Phage und verwendete ihn ursprünglich für Viren, die Bakterien infizieren. Phagen sind die größte derzeit bekannte biologische Gruppe. In der Literatur wird die auf der Welt vorhandene Zahl von Phagenpartikeln auf etwa 10^{31} geschätzt [3, 4]. Diese Menge übersteigt die Zahl der gegenwärtig im Universum bekannten Sterne. Das internationale Komitee für die Taxonomie der Viren (ICTV) ordnet diese riesige Zahl von Phagen in taxonomische Gruppen [5]. Derzeit umfasst die Speziesliste des ICTV 5.560 Virenspezies in 153 Familien und 16 Ordnungen.

Kein Urphage, kein Stammbaum

All das wirkt für biologisch Interessierte vertraut, ist es doch in der Begrifflichkeit an die Taxonomie der zellulären Organismen angelehnt. Doch es gibt zwei wesentliche Unterschiede aufgrund des Umstands, dass Phagen keine monophyletische Gruppe sind: Es gibt keinen gemeinsamen Vorfahren oder „Urphagen" für die Gesamtheit der Phagen und somit auch keinen Stammbaum, der alle Phagen umfasst. Und es gibt kein Gen, das in allen bekannten Phagengenomen vorhanden ist, und daher keine einheitliche sequenzbasierte Phylogenie. Bei den zellulären Organismen erstellt man sequenzbasierte Stammbäume auf Basis phylogenetischer Marker. Taxonomen verwenden hierfür das 16S-rRNA-Gen bei Prokaryoten [6] bzw. das 18S-rRNA-Gen bei höheren Organismen.

Die klassische Phagentaxonomie basiert auf einer Kombination deskriptiver Eigenschaften, wie Form und Größe der Phagenpartikel, der Organisation und Natur des genetischen Materials (Einzelstrang- oder Doppelstrang-Genom, RNA oder DNA) sowie des infizierten Wirts. Seit Beginn der Revolution durch die Sequenzierung der nächsten Generation (*Next Generation Sequencing*, NGS) – insbesondere seit deren Anwendung auf direkt aus der Umwelt isolierte DNA, die Metagenome – wurden in großem Umfang Sequenzdaten erstellt und öffentlich verfügbar gemacht. Darin befinden sich Tausende bisher unbekannte Phagengenome. Allerdings bestehen diese Daten nur aus Basenabfolgen. Daten, aus denen sich die klassische Phagentaxonomie ableiten lässt, sind für Phagengenomsequenzen nicht verfügbar. Entsprechend verfassten Peter Simmonds *et al.* im Jahr 2017 ein *Consensus Statement*, in dem der Bedarf an neuen bioinformatischen Werkzeugen festgestellt wurde, die es ermöglichen, Phagengenome ausschließlich auf Basis von Sequenzeigenschaften in das System der ICTV-Taxonomie zu integrieren [7].

Gigantische Phagen mit großen Genomen

Eine besondere Eigenschaft von Phagengenomen ist die Variabilität in Bezug auf das genetische Material, die Morphologie und die Genomgröße (**Tab. 1**). Die Mitglieder der Ordnung Caudovirales, die 1.320 von 5.560 Spezies aller vom ICTV akzeptierten Arten stellen, verfügen über Genomgrößen zwischen 35 und 200 Kilobasenpaaren (kbp). Typische Spezies der Inoviridae

(zu denen unter anderen der CTX-Phage aus *Vibrio cholera* und der M13-Phage aus *Escherichia coli* gehören) besitzen typische Genomgrößen von ca. 6 kbp mit elf Genen. Darüber hinaus sind *giant phages* bekannt mit Genomgrößen von 300 kbp bis 1,2 Megabasenpaaren, die bis zu 900 Gene codieren [8, 9]. Zum Vergleich: Das synthetische Bakterium JVCI-Syn3.0 verfügt über ein 531.490 Basenpaare großes Genom, das einen vollständig lebensfähigem Organismus codiert [10]. Im Gegensatz zu den zellulären Organismen teilen solche auch „Girus" (abgeleitet von *giant virus*) genannten DNA-Viren aber nur sehr wenige orthologe Gene, die als phylogenetische Marker verwendet werden können [11]. Eine taxonomische Einordnung solch unterschiedlich großer Genome auf Basis einzelner phylogenetischer Markermoleküle analog zur 16S-rRNA-Phylogenie macht für eine phylogenetische Einordnung nur begrenzt Sinn.

Gleichwohl können neue Phagengenome mithilfe von Sequenzvergleichen in die Phylogenie der Phagen eingeordnet werden. Auch auf die Phagen wirken die Kräfte der Evolution: Es gibt genetisches Material, das von Vorfahren an Nachkommen weitergeben wird. Zudem passen sich Gruppen von verwandten Phagengenomen durch adaptive Mutationen an sich ändernde Umweltbedingungen an. Innerhalb solcher Gruppen sind die einzelnen Mitglieder monophyletisch; somit lassen sich die Methoden aus der Phylogenie der zellulären Organismen anwenden, wie Meier-Kolthoff und Göker mit dem Programm VICTOR nachgewiesen haben [12]. Wir konnten zeigen, dass eine Kombination aus der Anzahl orthologer Gene unter Beachtung der Sequenzähnlichkeit auf Proteinebene ausreicht, um Mitglieder aus vier Phagenfamilien korrekt in ihre ICTV-Taxonomie einzuordnen (**Abb. 1**, [13]).

Phagenfamilien durch automatisierte Sequenzierung

Eine vergleichende Analyse von Phagengenomen aus den Familien Podoviridae, Myoviridae und Siphoviridae zusammen mit Genomen der Inoviridae zeigte eine eindeutige Zuordnung jedes Genoms zur korrekten Familie. Unter den getesteten Bedingungen erfolgte – entsprechend dem in Simmonds *et al.* [7] festgestellten Bedarf – eine rein sequenzbasierte Zuordnung von Phagengenomen zu den korrekten ICTV-Familien. In Anbetracht der schieren Menge an öffentlich zugänglichen Sequenzdaten kann dies nur mit einer automatisierten Methode erfolgen. Derzeit in Entwicklung befindliche bioinformatische Werkzeuge verwenden unter anderem die oben beschriebenen Eigenschaften, um mit Methoden des maschinellen Lernens Programme für die Identifizierung von neuen Mitgliedern von Phagenspezies aus Metagenomen zu ermöglichen.

Phagengenome lassen sich über die Anzahl und Qualität von Proteinvergleichen mit HMMs (*Hidden Markov Models*) aus familienspezifischen Proteinen eindeutig ihrer taxonomischen Familie zuordnen. Somit ist gezeigt, dass trotz der Abwesenheit eines gemeinsamen phylogenetischen Markers eine rein sequenzbasierte Taxonomie möglich ist. Um diese Methode der Taxonomie im Zuge der durch die Metagenomik explosiv wachsenden Zahl taxonomisch nicht klassifizierter Genome anwenden zu können, kann die beschriebene Methode in automatisierte Programme integriert werden. ■

Literatur

[1] Twort FW (1915) An investigation on the nature of ultra-microscopic viruses. Lancet 186:1241–1243
[2] D'Herelle F (1917) Sur un microbe invisible antagoniste des bacilles dysentériques. CR Acad Sci Paris 165:373–375
[3] Chow C-ET, Suttle CA (2015) Biogeography of viruses in the sea. Annu Rev Virol 2:41–66
[4] Breitbart M, Bonnain C, Malki K et al. (2018) Phage puppet masters of the marine microbial realm. Nat Microbiol 3:754–766
[5] Krupovic M, Dutilh BE, Adriaenssens EM et al. (2016) Taxonomy of prokaryotic viruses: update from the ICTV bacterial and archaeal viruses subcommittee. Arch Virol 161:1095–1099
[6] Nielsen HB, Almeida M, Juncker AS et al. (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol 32:822–828
[7] Simmonds P, Adams MJ, Benk M et al. (2017) Consensus statement: virus taxonomy in the age of metagenomics. Nat Rev Microbiol 15:161–168
[8] Yuan Y, Gao M (2017) Jumbo bacteriophages: an overview. Front Microbiol 8:1–9
[9] van Etten JL, Lane LC, Dunigan DD (2010) DNA viruses; the really big ones (giruses). Annu Rev Microbiol 64:83–99
[10] Glass JI, Merryman C, Wise KS et al. (2019) Minimal cells – real and imagined. Cold Spring Harb Perspect Biol 9, doi: 10.1101/cshperspect.a023861
[11] Gallot-Lavallée L, Blanc G, Claverie J-M (2017) Comparative genomics of *Chrysochromulina ericina* virus and other microalga-infecting large DNA viruses highlights their intricate evolutionary relationship with the established Mimiviridae family. J Virol 91, doi: 10.1128/JVI.00230-17
[12] Meier-Kolthoff JP, Göker M (2017) VICTOR : genome-based phylogeny and classification of prokaryotic viruses. Bioinformatics 33:3396–3404
[13] Chibani CM, Farr A, Klama S et al. (2019) Classifying the unclassified: a phage classification method. Viruses 11:195

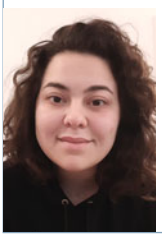
Funding: Open Access funding provided by Projekt DEAL.

Korrespondenzadressen:
Dr. Heiko Liesegang
Abteilung für genomische und angewandte Mikrobiologie
Institut für Mikrobiologie und Genetik
Universität Göttingen
Grisebachstraße 8
D-37077 Göttingen
hlieseg@gwdg.de

Dr. Cynthia Maria Chibani
Institut für Allgemeine Mikrobiologie
Christian-Albrechts-Universität zu Kiel
Am Botanischen Garten 1–9, R. 117
D-24118 Kiel
cchibani@ifam.uni-kiel.de

AUTOREN

Heiko Liesegang
Jahrgang **1963**. **1983–1990** Biologiestudium an der Universität Göttingen; dort **1990–1994** Promotion in Mikrobiologie bei Prof. Dr. H. G. Schlegel. **1994–2001** selbstständige Tätigkeit als EDV-Dozent und IT-Entwickler. **2001–2008** wissenschaftlicher Mitarbeiter im Laboratorium für Genomanalyse der Universität Göttingen. Seit **2008** Leiter einer Arbeitsgruppe für *Bacillus*-Genomik, industrielle Genomik und Bioinformatik in der Abteilung für Genomische und Angewandte Mikrobiologie von Prof. Dr. R. Daniel, Universität Göttingen.



Cynthia Maria Chibani
Jahrgang **1991**, geboren in Quebec, Kanada. **2009–2012** Biologiestudium an der Lebanese-American University – LAU, Byblos, Lebanon. **2013–2015** Masterstudium in Mikrobiologie und Biochemie an der Universität Göttingen. **2016–2019** Dort Promotion in Mikrobiologie bei Dr. H. Liesegang und Prof. Dr. R. Daniel. Seit **Juli 2019** wissenschaftliche Mitarbeiterin am Institut für Allgemeine Mikrobiologie in der Abteilung von Prof. Dr. R. Schmitz-Streit.