RUHR-UNIVERSITÄT BOCHUM

RUB

RUHR-UNIVERSITÄT BOCHUM

Ausarbeitung zum Thema "Mapping glycoprotein structure reveals Flaviviridae evolutionary history"

Tobias Polley

Wissenschaftliche Arbeit – 20. Dezember 2024
Lehrstuhl der Bioinformatik

Supervisor: Prof. Dr. Axel Mosig
Advisor: Prof. Dr. Martin Eisenacher

Eidesstattliche Erklärung

Ich erkläre, dass ich keine Arbeit in gleicher oder ähnlicher Fassung bereits für eine andere Prüfung an der Ruhr-Universität Bochum oder einer anderen Hochschule eingereicht habe.

Ich versichere, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Die Stellen, die anderen Quellen dem Wortlaut oder dem Sinn nach entnommen sind, habe ich unter Angabe der Quellen kenntlich gemacht. Dies gilt sinngemäß auch für verwendete Zeichnungen, Skizzen, bildliche Darstellungen und dergleichen.

Ich versichere auch, dass die von mir eingereichte schriftliche Version mit der digitalen Version übereinstimmt. Ich erkläre mich damit einverstanden, dass die digitale Version dieser Arbeit zwecks Plagiatsprüfung verwendet wird.

Date	Tobias Polley

Inhaltsverzeichnis

1 Einleitung

1.1 Vorstellung der Flaviviridae-Virusfamilie

Die Flaviviridae stellen eine der größten und vielfältigsten Familien von human- und tierpathogenen RNA-Viren dar [27]. Zu den wichtigsten klinischen Vertretern gehören Erreger wie das Dengue-Virus, Gelbfieber-Virus, Zika-Virus und das Hepatitis-C-Virus [16]. Diese Virenfamilie unterteilt sich in die vier etablierten Gattungen: *Flavivirus*, *Pestivirus*, *Pegivirus* und *Hepacivirus* [27]. Die jüngere Entdeckung von Jingmenviren und sogenannten Large Genome Flaviviruses (LGFs) hat die evolutionäre Vielfalt der Flaviviridae zusätzlich erweitert und verdeutlicht deren dynamische Anpassungsfähigkeit [26]. Die Identifikation von flavivirid Sequenzen in marinen Wirbellosen und basalen Wirbeltierlinien deutet darauf hin, dass die Evolution der Flaviviridae möglicherweise der Evolution der Metazoa durch Virus-Wirt-Kodivergenz über einen Zeitraum von Hunderten von Millionen Jahren folgt.

Glycoproteine spielen innerhalb der Flaviviridae eine Schlüsselrolle, da sie entscheidend den Eintritt des Virus in Zielzellen vermitteln und somit Wirtsspezifität und Pathogenität beeinflussen [20].

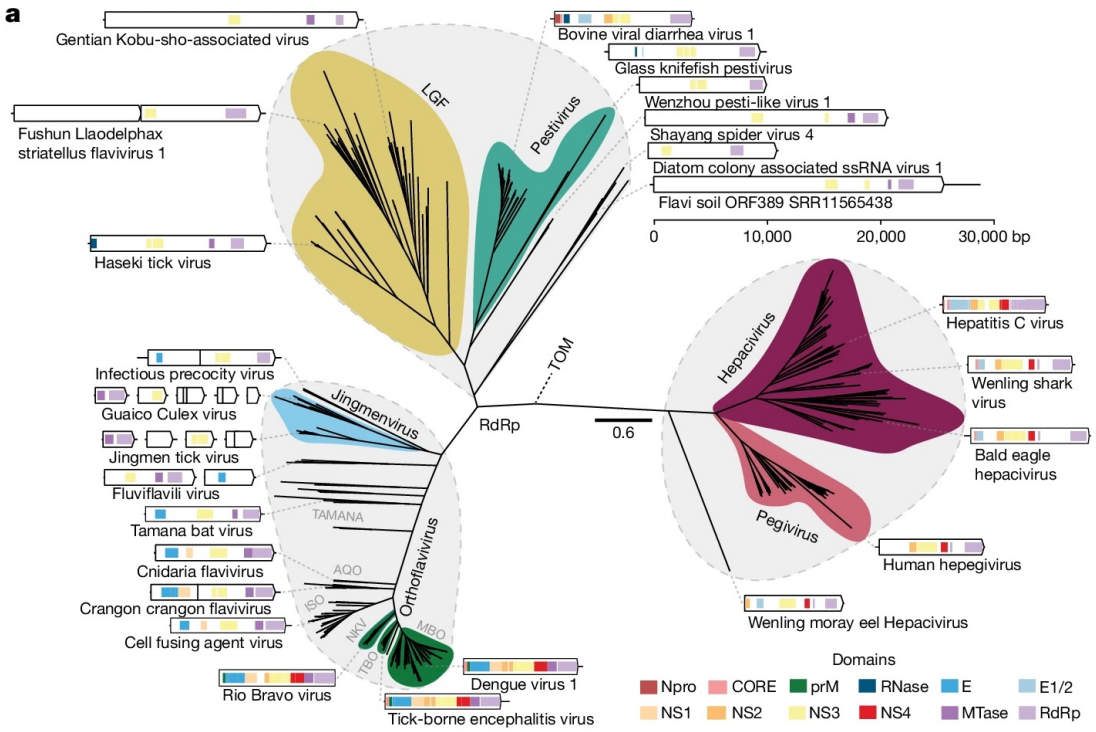

Abbildung 1.1: Protein Foldome der Flaviviridae-Familie.

Die Grafik zeigt die Vielfalt der Flaviviridae-Gattungen, darunter das Orthoflavivirus, Jingmenvirus, LGF und Pestivirus-ähnliche Viren. Die Fusion-Loop-Region ist ein struktureller Bestandteil, der eine zentrale Rolle bei der Membranfusion während des Viruseintritts in die Wirtszelle

spielt. Die abgebildeten Schleifen sind farblich kodiert, um die unterschiedlichen Gattungen zu repräsentieren und die strukturellen Gemeinsamkeiten sowie Unterschiede hervorzuheben. Diese Analyse ist essenziell für das Verständnis der evolutionären Beziehungen und funktionellen Diversifikation innerhalb der Flaviviridae-Familie. [17]

1.2 Bedeutung der Glycoprotein-Struktur für Evolution und Pathogenese

Die Glycoproteine der Flaviviridae erfüllen zentrale Funktionen in der Virusbiologie, insbesondere durch ihre Rolle bei der Membranfusion und der Interaktion mit dem Wirtsimmunsystem [8]. Während das E-Glycoprotein der *Flavivirus*-Gattungen als prototypisches Klasse-II-Fusionsprotein gut charakterisiert ist [12], weisen die Glycoproteine der Hepaciviren und Pegiviren, insbesondere die E1/E2-Komplexe, einzigartige strukturelle Eigenschaften auf. Diese könnten auf alternative Mechanismen der Membranfusion hindeuten und eine besondere evolutionäre Anpassung darstellen [13].

Mutationen, Rekombinationen und horizontaler Gentransfer haben entscheidend zur Diversifikation der Glycoprotein-Strukturen beigetragen und somit die Anpassungsfähigkeit der Flaviviridae an unterschiedliche ökologische Nischen gefördert [32]. Die Untersuchung dieser strukturellen Merkmale bietet daher wertvolle Einblicke in die evolutionären Mechanismen, die zur Spezialisierung und Pathogenität der Viren führen. Besonders konservierte Regionen wie die hydrophobe Fusion-Loop-Region und transmembrane Domänen spielen eine zentrale Rolle in der Funktion der Glycoproteine und unterliegen vermutlich starkem Selektionsdruck [19].

1.3 Zielsetzung und Vorgehensweise

Das Ziel dieser Arbeit besteht darin, die evolutionäre Geschichte und die strukturellen Eigenschaften der Glycoproteine der Flaviviridae systematisch zu untersuchen. Aufbauend auf der Arbeit von *Mifsud et al.* [17] werden bioinformatische Methoden eingesetzt, um die Phylogenie, Proteinstrukturvorhersage und strukturelle Homologien dieser Schlüsselproteine zu analysieren. Hierbei stehen drei zentrale Fragestellungen im Fokus:

Erstens sollen die strukturellen Gemeinsamkeiten und Unterschiede der Glycoproteine zwischen den Gattungen herausgearbeitet werden. Zweitens wird untersucht, welche evolutionären Mechanismen zur Diversifikation dieser Glycoproteine beigetragen haben. Drittens sollen die funktionellen Hinweise, die sich aus konservierten und variablen Strukturmotiven ableiten lassen, analysiert werden.

Die Arbeit gliedert sich wie folgt: Kapitel 2 beschreibt die eingesetzten bioinformatischen Verfahren zur Erstellung von Phylogenien, zur Proteinstrukturvorhersage und zur Homologiesuche. In Kapitel 3 werden die Ergebnisse zu den phylogenetischen Beziehungen und Proteinstrukturen vorgestellt. Kapitel 4 interpretiert diese Ergebnisse hinsichtlich ihrer evolutionären und funktionellen Bedeutung sowie methodischer Limitationen. Abschließend fasst Kapitel 5 die zentralen Erkenntnisse zusammen und gibt einen Ausblick auf zukünftige Forschungsansätze.

2 Methoden

2.1 Phylogenetische Analyse

Für die phylogenetische Untersuchung der Flaviviridae-Familie wurden 458 vollständige Genomsequenzen aus öffentlichen Datenbanken wie GenBank extrahiert. [17]. Die Auswahl der Genomsequenzen erfolgte auf Basis von Vollständigkeit, Annotation und Homologie, um fehlerhafte oder unvollständige Daten auszuschließen.

Als phylogenetischer Marker diente das Nichtstrukturelles Protein 5 (NS5)-Gen, welches die RNA-abhängige RNA-Polymerase (RNA-abhängige RNA-Polymerase (RdRp)) kodiert. Dieses Gen ist aufgrund seiner universellen Verbreitung und hohen Konservierung innerhalb der Flaviviridae ideal geeignet, um evolutionäre Beziehungen aufzuzeigen [11]. Die RdRp ist zentral für die virale RNA-Replikation und steht daher unter starkem Selektionsdruck, was ihre Sequenzstabilität über verschiedene Gattungen hinweg erklärt. Gleichzeitig enthält das Gen auch hinreichend variable Regionen, um divergente evolutionäre Linien zu identifizieren [21].

Die phylogenetische Analyse basiert auf einem Multiple Sequence Alignment (MSA), welches mithilfe des Programms Multiple Alignment using Fast Fourier Transform (MAFFT) erstellt wurde [10]. MAFFT kombiniert Fourier-Transformationen mit den Alignments, um homologe Regionen zu identifizieren. Diese Methode liefert robuste Alignments, selbst bei hochdivergenten Sequenzen. Die mathematische Grundlage der Fourier-Transformation, die für die Erkennung konservierter Sequenzmotive eingesetzt wird, ist im Anhang B.1 beschrieben. Alternativmodelle wie Clustal Omega oder MUSCLE sind hier nicht passend, aufgrund ihrer geringeren Sensitivität bei langen und hochvariablen Sequenzen [4].

Das resultierende MSA wurde zur Rekonstruktion eines phylogenetischen Baumes herangezogen. Hierbei wurde die Maximum-Likelihood-Methode angewandt, die eine probabilistische Modellierung der evolutionären Prozesse erlaubt [6]. Diese Methode optimiert die Wahrscheinlichkeit der beobachteten Daten unter Berücksichtigung der zugrunde liegenden Topologie des phylogenetischen Baumes und eines Substitutionsmodells. Für diese Arbeit wurde das General Time Reversible-Modell mit invarianten Positionen und gamma-verteilter Rate-Heterogenität (General Time Reversible Model with Invariant Sites and Gamma Distribution (GTR+I+G)) gewählt, da es eine flexible Modellierung von Nukleotidsubstitutionen ermöglicht und besonders für komplexe phylogenetische Beziehungen geeignet ist [29].

Die mathematische Formulierung des GTR+I+G-Modells, einschließlich der Parameter für Substitutionsraten, stationäre Wahrscheinlichkeiten und die Gamma-Verteilung, ist im Anhang B.2 näher erläutert. Dieses Modell berücksichtigt die Heterogenität der Substitutionsraten zwischen verschiedenen Sequenzpositionen, wodurch die Genauigkeit der phylogenetischen Topologie verbessert wird [33].

Um die Robustheit der rekonstruierten Topologie zu gewährleisten, wurde ein Bootstrapping mit 1.000 Wiederholungen durchgeführt [5]. Bootstrapping evaluiert die Stabilität der Baumknoten, indem es zufällige Resampling-Techniken auf das ursprüngliche Alignment anwendet und die resultierenden Bäume miteinander vergleicht. Hohe Bootstrap-Werte (>90 %) bestätigen die Zuverlässigkeit der Hauptknoten und ermöglichen eine fundierte Klassifizierung der

Hauptkladen innerhalb der Flaviviridae. Die spezifische Berechnung der Bootstrap-Werte sowie ihre Bedeutung für die Interpretation der phylogenetischen Ergebnisse sind im Anhang B.2 beschrieben.

Der endgültige phylogenetische Baum wurde mithilfe von Programmen wie FigTree und iTOL visualisiert, um die Hauptkladen und deren evolutionäre Beziehungen zu verdeutlichen [14]. Der Baum zeigt eine klare Aufteilung in drei Hauptkladen: *Flavivirus*, *Pegivirus/Hepacivirus* und *Pestivirus/Jingmenvirus/LGFs*, die durch signifikante Unterschiede in Sequenz- und Strukturmerkmalen gekennzeichnet sind (s.1.1).

2.2 Proteinstrukturvorhersage: ColabFold und ESMFold

Die dreidimensionalen Strukturen der Glycoproteine der Flaviviridae wurden mithilfe von Collaborative Protein Folding (ColabFold) und Evolutionary Scale Modeling for Protein Folding (ESMFold) vorhergesagt.

ColabFold, eine Jupyter-Notebook Implementierung von AlphaFold2. Hierbei werden Multiple Sequence Alignments (MSAs) mit Transformer-Netzwerken kombiniert, um präzise Strukturvorhersagen zu liefern [18]. Zunächst werden homologe Sequenzen aus großen Datenbanken wie UniProt (UniRef90) und Big Fantastic Database (BFD) identifiziert und in einem MSA zusammengefasst. Die daraus extrahierten evolutionären Informationen werden im Evoformer-Modul verarbeitet, um langreichweitige Wechselwirkungen zwischen Aminosäuren zu modellieren. Abschließend erfolgt die Vorhersage der Proteinstruktur, deren Qualität mithilfe der Predicted Local Distance Difference Test (pLDDT)-Werte bewertet wird. Die mathematische Formulierung der pLDDT-Werte und die Berechnung der Root-Mean-Square Deviation (Root-Mean-Square Deviation (RMSD)) zur Bewertung der Strukturen sind im Anhang B.3 näher beschrieben.

Im Gegensatz dazu verwendet ESMFold einen sprachmodellbasierten Ansatz, der keine MSAs benötigt. Hierbei wird die Proteinsequenz direkt durch ein Transformer-Sprachmodell analysiert, das auf Millionen von Sequenzen trainiert wurde [15]. Besonders bei hochdivergenten Proteinen mit wenigen homologen Sequenzen liefert ESMFold zuverlässige Ergebnisse, indem es die Sequenzinformationen kontextabhängig interpretiert.

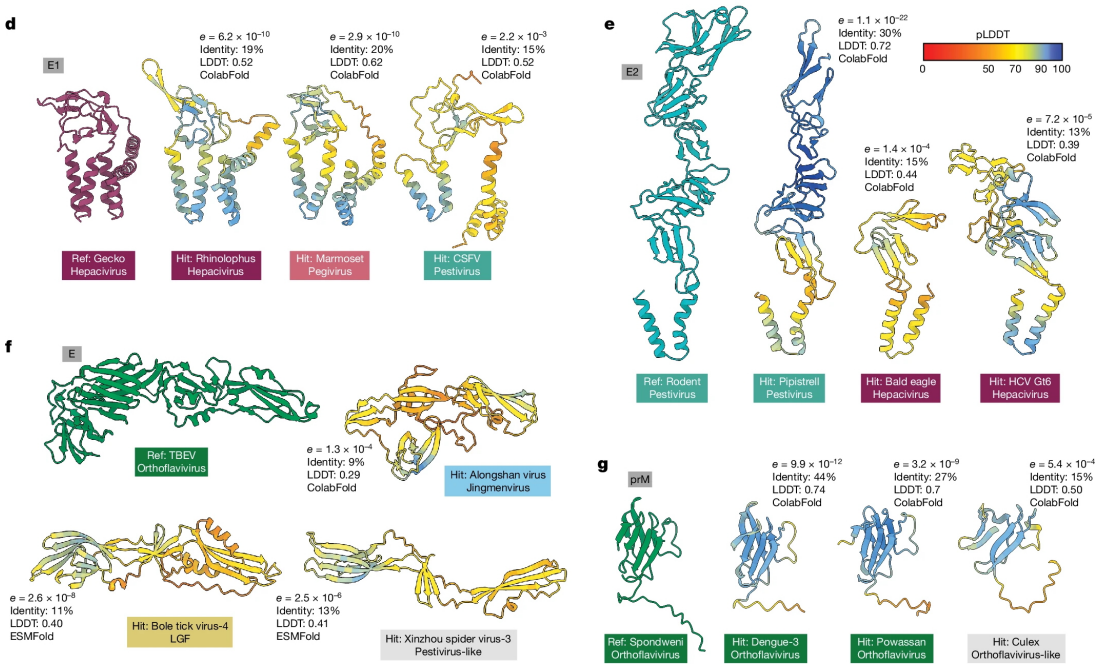

Abbildung 2.1: Illustration der pLDDT-Werte in Korrelation mit den MSA-Alignments

Die Abbildung zeigt die Korrelation zwischen den pLDDT-Werten und den MSA-Alignments. Die Farbskala repräsentiert die Zuverlässigkeit der Strukturvorhersagen, wobei hohe pLDDT-Werte (blau) auf eine hohe Übereinstimmung mit den Alignments hinweisen. Die Analyse der pLDDT-Werte ermöglicht eine präzise Bewertung der Strukturqualität und Identifikation von konservierten Regionen innerhalb der Glycoproteine.

2.3 Homologiesuche: Anwendung von Foldseek

Zur Identifizierung struktureller Homologien zwischen den vorhergesagten Glycoprotein-Strukturen wurde das Programm Fast Protein Structure Search Tool (Foldseek) eingesetzt [30]. Foldseek ermöglicht einen schnellen Vergleich von Proteinstrukturen, indem es deren dreidimensionale Koordinaten in vereinfachte Merkmalsrepräsentationen umwandelt. Diese Methode ist besonders effizient, um strukturelle Ähnlichkeiten auch bei hoher Sequenzdivergenz zu erkennen, was klassische sequenzbasierte Methoden wie BLAST nicht leisten können [1].

Die vorhergesagten Glycoprotein-Strukturen aus ColabFold und ESMFold wurden mit bekannten Strukturen aus der Protein Data Bank (Protein Data Bank (PDB)) verglichen. Die Übereinstimmung der Modelle wurde durch Alignment-Scores und RMSD-Werte quantifiziert, deren mathematische Grundlagen im Anhang B.3 erläutert werden. Konservierte Faltungsmuster und potenzielle funktionelle Gemeinsamkeiten, wie die Klasse-II-Fusionsproteinfaltung, wurden identifiziert.

Die Ergebnisse der Homologiesuche wurden anschließend mit den phylogenetischen Analysen kombiniert, um divergente und konvergente evolutionäre Anpassungen innerhalb der Flaviviridae zu identifizieren. Auf diese Weise konnten strukturelle Gemeinsamkeiten zwischen verschiedenen Gattungen trotz erheblicher Sequenzunterschiede nachgewiesen werden.

3 Ergebnisse

3.1 Phylogenetische Strukturen und Klassifizierung

Die phylogenetische Analyse der Flaviviridae-Familie, basierend auf den NS5-Sequenzen von 458 vollständigen Virengenomen, führte zu einer klaren Aufteilung in drei Hauptkladen. Die Maximum-Likelihood-Methode wurde zur Rekonstruktion des phylogenetischen Baumes verwendet, wobei das Substitutionsmodell GTR+I+G zum Einsatz kam. Die resultierende Topologie wurde durch hohe Bootstrap-Werte (>90%) unterstützt und ermöglicht eine robuste Klassifizierung [17].

Die erste Hauptklade umfasst klassische Vertreter der Gattung *Flavivirus*, zu denen das Dengue-Virus, Zika-Virus und Gelbfieber-Virus gehören. Diese Gruppe zeigt eine bemerkenswerte Konservierung des NS5-Gens, was auf eine enge evolutionäre Verwandtschaft hindeutet. Die Glycoprotein-Strukturen, insbesondere das E-Glycoprotein, sind stark konserviert und spielen eine zentrale Rolle beim Eintritt in Wirtszellen [24].

In der zweiten Hauptklade wurden die Gattungen *Pegivirus* und *Hepacivirus* zusammengefasst. Hier zeigt sich eine größere genetische Diversität, insbesondere in den Glycoprotein-Genen. Die strukturelle Variabilität der E1/E2-Glycoproteine deutet auf spezifische Mechanismen der Wirtsspezifität und des Viruseintritts hin [31].

Die dritte Hauptklade vereint die Gattungen *Pestivirus*, *Jingmenvirus* und die sogenannten Large Genome Flaviviruses (LGFs). Diese Gruppe weist stark divergente Sequenzen und eine hohe strukturelle Vielfalt in den Glycoproteinen auf. Auffällig sind hier Anpassungen in transmembranen Domänen und möglichen Rezeptorbindungsstellen, die auf spezifische Wirt-Interaktionen hinweisen [28].

3.2 Glycoprotein-Divergenz: Unterschiede zwischen Gattungen

Die Vorhersage der Glycoprotein-Strukturen mithilfe von ColabFold und ESMFold zeigte sowohl konservierte als auch variable Merkmale zwischen den Hauptkladen.

Bei den *Flaviviren* erwies sich das E-Glycoprotein als stark konserviert, insbesondere in der hydrophoben Fusion-Loop-Region und den transmembranen Domänen. Die typische Klasse-II-Faltungsarchitektur mit drei Domänen aus β-Faltblättern wurde bestätigt und unterstreicht die essenzielle Funktion dieser Region [19].

Im Gegensatz dazu zeigten die Glycoproteine der *Hepaciviren* und *Pegiviren*, die aus E1/E2-Komplexen bestehen, signifikante strukturelle Unterschiede. Vor allem die Oberflächenexpositionen potenzieller Rezeptorbindungsstellen variieren stark und könnten die breite Wirtsspezifität und unterschiedliche Pathogenitätsprofile dieser Gattungen erklären [13].

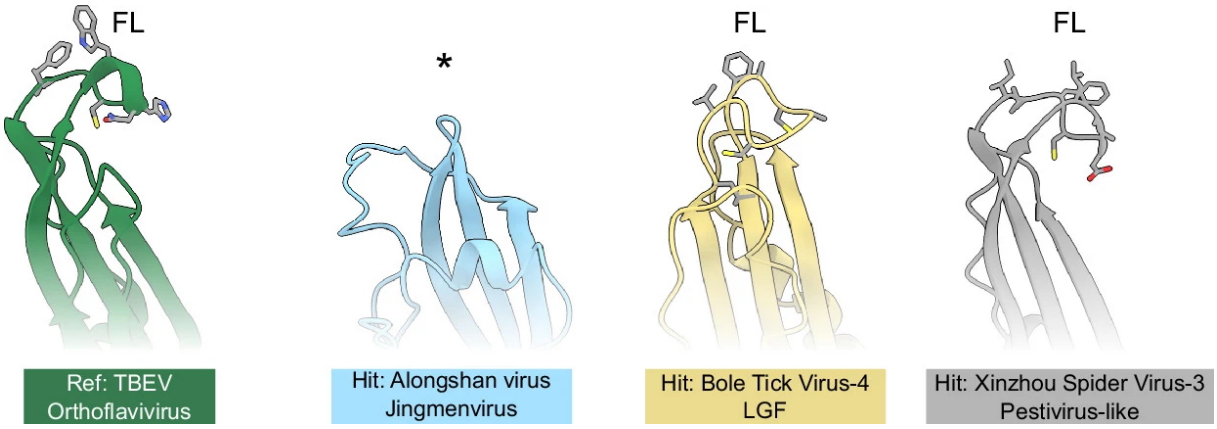

Abbildung 3.1: Strukturelle Unterschiede der Fusion-Loop-Regionen

Die größte strukturelle Diversität wurde in den Glycoproteinen der *Pestiviren* und *Jingmenviren* beobachtet. Unterschiede in der Anzahl und Position der transmembranen Domänen sowie Variationen in den extrazellulären Regionen deuten auf funktionelle Anpassungen hin, die spezifische Interaktionen mit Wirtsorganismen ermöglichen [28].

Trotz der beobachteten Variabilität wurden konservierte Motive wie die hydrophoben Fusion-Loops und transmembranen Segmente identifiziert, die für die grundlegende Funktion der Glycoproteine entscheidend sind und unter starkem Selektionsdruck stehen [19].

3.3 Spezifische Strukturmerkmale (Fusion-Loop und transmembrane Regionen)

Die detaillierte Analyse der Fusion-Loop-Regionen und transmembranen Segmente der Glycoproteine lieferte wichtige Einblicke in deren funktionelle Rollen. Die Fusion-Loops bestehen aus konservierten hydrophoben Aminosäureresten, die essenziell für die Membranfusion sind. Bei den *Flaviviren* zeigt die Fusion-Loop eine charakteristische Schleifenstruktur, die von Glycinresten flankiert wird und Flexibilität unter sauren pH-Bedingungen ermöglicht [24]. Im Gegensatz dazu deuten Unterschiede in Sequenz und Sekundärstruktur bei *Hepaciviren* und *Pegiviren* auf alternative Mechanismen der Membranfusion hin [13].

Die transmembranen Regionen variieren stark zwischen den Gattungen. Während *Flaviviren* typischerweise nur eine einzelne transmembrane Domäne am C-Terminus des E-Glycoproteins besitzen, weisen *Pestiviren* und *Hepaciviren* mehrere transmembrane Segmente auf. Diese Variationen könnten die Organisation der Glycoproteine in der viralen Membran beeinflussen und somit die Virusassemblierung und -freisetzung modulieren [23]. Zusätzlich wurden Signalpeptide und Ankersequenzen identifiziert, die eine wichtige Rolle für die Lokalisierung und Stabilität der Proteine spielen [2].

4 Diskussion

4.1 Evolutionäre Bedeutung der Glycoprotein-Divergenz

Die Ergebnisse dieser Studie betonen die zentrale Rolle der Glycoproteine in der Evolution und Anpassung der Flaviviridae-Familie. Die beobachtete strukturelle Diversität reflektiert eine dynamische Evolution, die durch Selektionsdruck, Wirtsspezifität und ökologische Anpassung geprägt ist [17].

Konservierte Regionen wie die hydrophoben Fusion-Loops und transmembranen Domänen zeigten eine bemerkenswerte Stabilität über verschiedene Gattungen hinweg. Dies unterstreicht ihre essentielle Funktion bei der Membranfusion und beim Eintritt des Virus in die Wirtszelle [19, 24]. Die evolutionäre Stabilität dieser Regionen ist vermutlich auf starken Selektionsdruck zurückzuführen, da selbst geringfügige Veränderungen in diesen Bereichen die Infektiosität und Replikationsfähigkeit der Viren erheblich beeinträchtigen könnten.

Gleichzeitig weisen die variablen Regionen der Glycoproteine, wie die Oberflächenexpositionen und Glycosylierungsstellen, eine hohe Diversität auf. Diese strukturellen Unterschiede ermöglichen es den Viren, sich an verschiedene Wirte anzupassen und das Immunsystem gezielt zu umgehen [13]. Insbesondere die strukturelle Variabilität der E1/E2-Komplexe bei Hepaciviren und Pegiviren könnte die Fähigkeit zur breiten Wirtsspezifität sowie zur Etablierung chronischer Infektionen erklären [31].

Die beobachteten Unterschiede zwischen den Gattungen lassen sich zudem auf evolutionäre Mechanismen wie Rekombinationsereignisse und horizontalen Gentransfer zurückführen. Durch den Austausch genetischen Materials zwischen verschiedenen Viruslinien oder den Erwerb neuer Gene konnten die Flaviviridae ihre Anpassungsfähigkeit an unterschiedliche Umweltbedingungen und Wirte verbessern [32].

4.2 Methodologische Limitierungen und Unsicherheiten

Trotz der robusten Ergebnisse und der breiten Anwendung bioinformatischer Methoden weist diese Studie einige methodische Einschränkungen auf. Die Genauigkeit der Strukturvorhersagen hängt stark von der Verfügbarkeit homologer Sequenzen und experimenteller Referenzdaten ab. Bei hochdivergenten Glycoproteinen, wie denen der Jingmenviren, könnte die Zuverlässigkeit der in silico-Modelle von ColabFold und ESMFold eingeschränkt sein [9].

Darüber hinaus beruhen die Ergebnisse ausschließlich auf Vorhersagen und Vergleichen. Experimentelle Validierung der vorhergesagten Strukturen durch Kryo-Elektronenmikroskopie oder Röntgenkristallographie ist notwendig, um die Modelle zu bestätigen und ihre funktionelle Relevanz zu überprüfen [3]. Eine weitere Limitation liegt in der Fokussierung auf ein einzelnes phylogenetisches Marker-Gen (NS5). Obwohl dieses Gen aufgrund seiner Konservierung geeignet ist, könnten komplexe evolutionäre Beziehungen durch eine multimarkergestützte Analyse präziser abgebildet werden [5].

4.3 Zukünftige Forschungsperspektiven

Die vorliegende Studie eröffnet mehrere wichtige Forschungsrichtungen, die das Verständnis der Evolution und Struktur-Funktion-Beziehungen der Glycoproteine der Flaviviridae weiter vertiefen können. Ein zentraler Aspekt zukünftiger Arbeiten sollte die experimentelle Validierung der vorhergesagten Glycoprotein-Strukturen sein. Die Anwendung von Kryo-Elektronenmikroskopie und Röntgenstrukturanalyse könnte die Zuverlässigkeit der bioinformatischen Modelle erhöhen und neue funktionelle Erkenntnisse zu den analysierten Regionen liefern.

Darüber hinaus sollten detaillierte Studien der variablen Oberflächenregionen und Glycosylierungsstellen erfolgen, um neue Angriffspunkte für antivirale Strategien zu identifizieren. Diese Regionen spielen eine zentrale Rolle bei der Wirtsspezifität und Immunumgehung und könnten zur Entwicklung gattungsspezifischer Therapien und Impfstoffe beitragen [7].

Die Erweiterung der phylogenetischen Analysen durch zusätzliche Marker und die Integration von Metagenomik-Daten würde eine präzisere Auflösung der evolutionären Beziehungen innerhalb der Flaviviridae ermöglichen. Dies könnte nicht nur zur Entdeckung neuer Viruslinien führen, sondern auch die Risiken potenzieller zoonotischer Übertragungen besser abschätzen lassen.

Darüber hinaus bietet die Weiterentwicklung bioinformatischer Methoden großes Potenzial. Die Kombination von maschinellen Lernverfahren mit experimentellen Daten könnte die Strukturvorhersage für schlecht charakterisierte Proteine verbessern und neue Wege zur Analyse funktioneller Proteindomänen eröffnen [25].

5 Schlussfolgerung

Die vorliegende Arbeit liefert neue Erkenntnisse zur evolutionären Geschichte und strukturellen Diversität der Glycoproteine innerhalb der Flaviviridae-Familie. Durch die Kombination von phylogenetischen Analysen, Proteinstrukturvorhersagen mittels ColabFold und ESMFold sowie strukturellen Homologiesuchen mit Foldseek konnten sowohl konservierte als auch variable Merkmale der Glycoproteine identifiziert und charakterisiert werden [17].

5.1 Wichtige Ergebnisse und Implikationen

Die phylogenetische Analyse unterteilte die Flaviviridae in drei Hauptkladen: *Flavivirus*, *Pegivirus/Hepacivirus* und *Pestivirus/Jingmenvirus/LGFs*. Diese Aufteilung korreliert eng mit den strukturellen Eigenschaften der Glycoproteine und verdeutlicht die evolutionäre Verknüpfung zwischen genetischer Sequenzkonservierung und funktioneller Diversifikation. Konservierte Regionen wie die hydrophoben Fusion-Loops und transmembranen Domänen wurden über alle Hauptkladen hinweg nachgewiesen und unterstreichen ihre essentielle Rolle bei der Membranfusion und dem Viruseintritt [19]. Aufgrund dieser funktionellen Konservierung bieten sie vielversprechende Zielstrukturen für breit wirksame antivirale Therapien und Impfstoffentwicklungen [7].

Gleichzeitig wurden in den Glycoproteinen signifikante strukturelle Variationen festgestellt, die adaptive Mechanismen widerspiegeln. Insbesondere die E1/E2-Komplexe der Hepaciviren und Pegiviren zeigten eine erhebliche Diversität in der räumlichen Anordnung der Oberflächenregionen, was auf eine mögliche Anpassung an unterschiedliche Wirte und Immunabwehrmechanismen hinweist [31, 13]. Diese Variationen eröffnen neue Ansätze für gattungsspezifische antivirale Strategien, die auf variablen Regionen wie den Glycosylierungsstellen basieren.

Methodisch demonstrierte die Arbeit die Leistungsfähigkeit moderner bioinformatischer Werkzeuge zur Strukturvorhersage und Homologieerkennung. Die Kombination von ColabFold und ESMFold lieferte detaillierte Einblicke in die dreidimensionalen Strukturen der Glycoproteine, offenbarte jedoch auch die Notwendigkeit experimenteller Validierung. Dies gilt insbesondere für hochdivergente Proteine, bei denen die Vorhersagegenauigkeit durch die begrenzte Verfügbarkeit homologer Referenzdaten eingeschränkt sein könnte.

5.2 Persönliche Kritik und Reflexion

Die zugrundeliegende Arbeit hebt hervor, wie interdisziplinäre Ansätze, insbesondere die Anwendung bioinformatischer Werkzeuge, einen entscheidenden Beitrag zum Verständnis der Flaviviridae-Familie leisten können. Dennoch gibt es wesentliche Aspekte, die einer kritischen Reflexion bedürfen.

Die Nutzung moderner Werkzeuge wie ColabFold, ESMFold und Foldseek zeigt, wie maschinelles Lernen die Proteinstrukturvorhersage und phylogenetische Analyse vorantreiben. Diese Technologien demonstrieren nicht nur eine hohe Effizienz und Präzision, sondern auch die Fähigkeit,

komplexe biologische Fragestellungen in einem breiten Maßstab zu adressieren. Besonders positiv hervorzuheben ist die konsistente Nutzung standardisierter bioinformatischer Pipelines, die sowohl Reproduzierbarkeit als auch Skalierbarkeit gewährleisten. Durch die stetige Weiterentwicklung dieser Methoden könnten zukünftige Studien noch tiefere Einblicke in die Struktur-Funktions-Beziehungen der Glycoproteine ermöglichen.

Trotz der Fortschritte in der Proteinstrukturvorhersage bleibt die Abhängigkeit von in silico Methoden eine zentrale Limitation. Die Vorhersagegenauigkeit ist bei hochdivergenten Glycoproteinen, wie sie in dieser Arbeit untersucht wurden, eingeschränkt, insbesondere wenn homologe Referenzstrukturen fehlen. Dies unterstreicht die Notwendigkeit, maschinelle Lernmodelle kontinuierlich durch experimentelle Daten zu validieren und zu verbessern. Dise ist eine Herausforderung und zugleich eine Chance, durch erweiterte Trainingsdatensätze die Vorhersagekraft zu steigern.

Diese Arbeit verdeutlicht eindrucksvoll, wie stark Methoden des maschinellen Lernens das Verständnis komplexer biologischer Phänomene erweitern können. Die Arbeit hat mir gezeigt, wie wichtig die Verbindung von technischen Kompetenzen mit domänenspezifischem Wissen ist, um interdisziplinäre Herausforderungen zu bewältigen. In Zukunft wird eine noch engere Zusammenarbeit zwischen Bioinformatikern, Virologen und Strukturbiologen entscheidend sein, um die vielfältigen Facetten der Virus-Familien zu entschlüsseln und neue Therapieansätze zu entwickeln.

5.3 Zukünftige Forschungsrichtungen

Basierend auf den Ergebnissen dieser Arbeit ergeben sich mehrere zentrale Forschungsrichtungen für die Zukunft. Ein vorrangiges Ziel sollte die experimentelle Validierung der vorhergesagten Glycoprotein-Strukturen sein. Methoden wie Kryo-Elektronenmikroskopie oder Röntgenkristallographie könnten die Präzision der in silico-Modelle bestätigen und zusätzliche Einblicke in die funktionelle Bedeutung der konservierten und variablen Regionen liefern [3].

Erweiterte phylogenetische Analysen, die zusätzliche genetische Marker einbeziehen, könnten die Auflösung der evolutionären Beziehungen weiter verbessern. Die Integration von Metagenomik-Daten bietet zudem die Möglichkeit, bisher unbekannte Viruslinien zu identifizieren und ein umfassenderes Verständnis der Diversität und Adaptionsmechanismen der Flaviviridae zu gewinnen. Dies wäre besonders wertvoll, um zoonotische Risiken und die Anpassung von Viren an neue Wirtsorganismen besser abschätzen zu können.

Ein weiterer Schwerpunkt zukünftiger Forschung sollte auf die therapeutische Anwendung der Erkenntnisse gelegt werden. Die konservierten Regionen, wie die Fusion-Loops und transmembranen Domänen, stellen ideale Angriffspunkte für breit wirksame antivirale Wirkstoffe und Impfstoffe dar. Gleichzeitig könnten gattungsspezifische strukturelle Unterschiede zur Entwicklung gezielter antiviraler Strategien genutzt werden, die spezifisch auf variablen Regionen der Glycoproteine basieren.

Schließlich eröffnet die Weiterentwicklung bioinformatischer Methoden vielversprechende Perspektiven. Die Kombination von maschinellem Lernen mit experimentellen Daten könnte die Vorhersagegenauigkeit für schlecht charakterisierte Proteine erheblich verbessern und neue Möglichkeiten zur Analyse funktioneller Proteindomänen bieten [25]. Verbesserungen in der Effizienz und Präzision dieser Modelle würden insbesondere bei hochdivergenten Glycoproteinen von großer Bedeutung sein.

Glossar

BFD Big Fantastic Database. 4

ColabFold Collaborative Protein Folding. 4–6, 8, 10

ESMFold Evolutionary Scale Modeling for Protein Folding. 4–6, 8, 10

Foldseek Fast Protein Structure Search Tool. 5, 10

GTR+I+G General Time Reversible Model with Invariant Sites and Gamma Distribution. 3, 6

MAFFT Multiple Alignment using Fast Fourier Transform. 3

MSA Multiple Sequence Alignment. 3–5, 13, 18

NS5 Nichtstrukturelles Protein 5. 3, 6

PDB Protein Data Bank. 5

pLDDT Predicted Local Distance Difference Test. 4

RdRp RNA-abhängige RNA-Polymerase. 3

RMSD Root-Mean-Square Deviation. 4, 5, 20

UniRef90 UniProt. 4

Abbildungsverzeichnis

A List of AI-Tools

Bei der Erstellung dieser Arbeit wurden mehrere KI-Tools eingesetzt, um den Forschungs- und Schreibprozess zu verbessern. Ich erkläre hiermit, dass ich alle von der KI generierten Vorschläge sorgfältig auf ihre Korrektheit überprüft habe. Die folgenden Tools waren dabei besonders hilfreich:

ChatGPT Wurde für das Brainstorming von Forschungsfragen und zur Umformulierung von Texten verwendet, um die Lesbarkeit zu verbessern.

Grammarly Wurde zum Korrekturlesen und zur Verbesserung der allgemeinen Schreibqualität eingesetzt.

DeepL Angewandt für die Übersetzung von Schlüsselpassagen aus englischen Quellen ins Deutsche.

Literatur

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers und D. J. Lipman. „Basic Local Alignment Search Tool“. In: *Journal of Molecular Biology* 215.3 (1990), S. 403–410. DOI: 10.1016/S0022-2836(05)80360-2.

[2] Haim Ashkenazy, Shlomit Abadi, Eric Martz, Ofer Chay, Itay Mayrose, Tal Pupko und Nir Ben-Tal. „ConSurf 2016: An Improved Methodology to Estimate and Visualize Evolutionary Conservation in Macromolecules“. In: *Nucleic Acids Research* 44.W1 (2016), W344–W350. DOI: 10.1093/nar/gkw408.

[3] Ewen Callaway. „Revolutionizing Structural Biology with Cryo-EM“. In: *Nature* 578.7794 (2020), S. 201–202. DOI: 10.1038/d41586-020-00341-9.

[4] Robert C. Edgar. „MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput“. In: *Nucleic Acids Research* 32.5 (März 2004), S. 1792–1797. ISSN: 0305-1048. DOI: 10.1093/nar/gkh340. (Besucht am 20. 12. 2024).

[5] Joseph Felsenstein. „Confidence Limits on Phylogenies: An Approach Using the Bootstrap“. In: *Evolution; international journal of organic evolution* 39.4 (1985), S. 783–791. DOI: 10.2307/2408678.

[6] Joseph Felsenstein. „Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach“. In: *Journal of Molecular Evolution* 17.6 (Nov. 1981), S. 368–376. ISSN: 1432-1432. DOI: 10.1007/BF01734359. (Besucht am 20. 12. 2024).

[7] Alfonso Fernandez, Ozlem Keskin und Attila Gursoy. „Engineering Antibody Therapeutics“. In: *Current Opinion in Structural Biology* 52 (2018), S. 15–22. DOI: 10.1016/j.sbi.2018.07.003.

[8] Franz X. Heinz und Karin Stiasny. „Flaviviruses and Their Antigenic Structure“. In: *Journal of Clinical Virology* 55.4 (2012), S. 289–295. DOI: 10.1016/j.jcv.2012.08.024.

[9] John Jumper, Richard Evans, Alexander Pritzel und et al. „Highly Accurate Protein Structure Prediction with AlphaFold“. In: *Nature* 596.7873 (2021), S. 583–589. DOI: 10.1038/s41586-021-03819-2.

[10] Kazutaka Katoh und Daron M. Standley. „MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability“. In: *Molecular Biology and Evolution* 30.4 (2013), S. 772–780. DOI: 10.1093/molbev/mst010.

[11] Eugene V. Koonin. „The Phylogeny of RNA-dependent RNA Polymerases of Positive-Strand RNA Viruses“. In: *Journal of General Virology* 72.9 (1991), S. 2197–2206. DOI: 10.1099/0022-1317-72-9-2197.

[12] Richard J. Kuhn, Wei Zhang, Michael G. Rossmann, Sergei V. Pletnev, Jeroen Corver, Edith Lenches, Christopher T. Jones, Suchetana Mukhopadhyay, Paul R. Chipman, Ellen G. Strauss, Timothy S. Baker und James H. Strauss. „Structure of Dengue Virus: Implications for Flavivirus Organization, Maturation, and Fusion“. In: *Cell* 108.5 (März 2002), S. 717–725. ISSN: 0092-8674. DOI: 10.1016/s0092-8674(02)00660-8.

[13] Marion Lavie und Jean Dubuisson. „Interplay between Hepatitis C Virus and Lipid Metabolism during Virus Entry and Assembly“. In: *Biochimie* 141 (2017), S. 62–69. DOI: 10.1016/j.biochi.2017.06.013.

[14] Ivica Letunic und Peer Bork. „Interactive Tree Of Life (iTOL) v5: An Online Tool for Phylogenetic Tree Display and Annotation“. In: *Nucleic Acids Research* 49.W1 (Juli 2021), W293–W296. ISSN: 1362-4962. DOI: 10.1093/nar/gkab301.

[15] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido und Alexander Rives. „Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model“. In: *Science* 379.6637 (März 2023), S. 1123–1130. DOI: 10.1126/science.ade2574. (Besucht am 01. 12. 2024).

[16] John S. Mackenzie, Duane J. Gubler und Lyle R. Petersen. „Emerging Flaviviruses: The Spread and Resurgence of Japanese Encephalitis, West Nile and Dengue Viruses“. In: *Nature Medicine* 10.12 Suppl (2004), S98–S109. DOI: 10.1038/nm1144.

[17] Jonathon C. O. Mifsud, Spyros Lytras, Michael R. Oliver, Kamilla Toon, Vincenzo A. Costa, Edward C. Holmes und Joe Grove. „Mapping Glycoprotein Structure Reveals Flaviviridae Evolutionary History“. In: *Nature* 633.8030 (Sep. 2024), S. 695–703. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07899-8. (Besucht am 01. 12. 2024).

[18] Markus Mirdita, Konstantin Schütze, Yoshitaka Moriwaki und et al. „ColabFold: Making Protein Folding Accessible to All“. In: *Nature Methods* 19.6 (2022), S. 679–682. DOI: 10.1038/s41592-022-01488-1.

[19] Yorgo Modis, Sadie Ogata, Don Clements und Stephen C. Harrison. „Structure of the Dengue Virus Envelope Protein after Membrane Fusion“. In: *Nature* 427.6972 (2004), S. 313–319. DOI: 10.1038/nature02165.

[20] Suchetana Mukhopadhyay, Richard J. Kuhn und Michael G. Rossmann. „A Structural Perspective of the Flavivirus Life Cycle“. In: *Nature Reviews Microbiology* 3.1 (2005), S. 13–22. DOI: 10.1038/nrmicro1067.

[21] Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler und Bui Quang Minh. „IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies“. In: *Molecular Biology and Evolution* 32.1 (2015), S. 268–274. DOI: 10.1093/molbev/msu300.

[22] Michael Pees, Volker Schmidt, Tibor Papp, Ákos Gellért, Maha Abbas, J. Matthias Starck, Annkatrin Neul und Rachel E. Marschang. „Three Genetically Distinct Ferlaviruses Have Varying Effects on Infected Corn Snakes (Pantherophis Guttatus)“. In: *PLOS ONE* 14.6 (Juni 2019), e0217164. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0217164. (Besucht am 20. 12. 2024).

[23] François Penin, Volker Brass, Nicole Appel, Stephanie Ramboarina, Roland Montserret, Damien Ficheux, Hubert E. Blum, Ralf Bartenschlager und Darius Moradpour. „Structure and Function of the Membrane Anchor Domain of Hepatitis C Virus Nonstructural Protein 5A“. In: *The Journal of Biological Chemistry* 279.39 (Sep. 2004), S. 40835–40843. ISSN: 0021-9258. DOI: 10.1074/jbc.M404761200.

[24] Félix A. Rey, Franz X. Heinz, Christiane Mandl, Christian Kunz und Stephen C. Harrison. „The Envelope Glycoprotein from Tick-Borne Encephalitis Virus at 2 Å Resolution“. In: *Nature* 375.6529 (1995), S. 291–298. DOI: 10.1038/375291a0.

[25] Andrew W. Senior, Richard Evans, John Jumper und et al. „Improved Protein Structure Prediction Using Potentials from Deep Learning“. In: *Nature* 577.7792 (2020), S. 706–710. DOI: 10.1038/s41586-019-1923-7.

[26] Mang Shi, Xian-Dan Lin, Nikos Vasilakis, Jun-Hua Tian, Ci-Xiu Li, Liang-Jun Chen, Gillian Eastwood, Xiu-Nian Diao, Ming-Hui Chen, Xiao Chen, Xin-Cheng Qin, Steven G. Widen, Thomas G. Wood, Robert B. Tesh, Jianguo Xu, Edward C. Holmes und Yong-Zhen Zhang. „Divergent Viruses Discovered in Arthropods and Vertebrates Revise the Evolutionary History of the Flaviviridae and Related Viruses“. In: *Journal of Virology* 90.2 (Dez. 2015), S. 659–669. ISSN: 0022-538X. DOI: 10.1128/JVI.02036-15. (Besucht am 01. 12. 2024).

[27] Peter Simmonds, Paul Becher, Michael S. Collett und et al. „ICTV Virus Taxonomy Profile: Flaviviridae“. In: *Journal of General Virology* 98.1 (2017), S. 2–3. DOI: 10.1099/jgv.0.000672.

[28] Niels Tautz, Birke A. Tews und Gunter Meyers. „The Molecular Biology of Pestiviruses“. In: *Advances in Virus Research* 93 (2015), S. 47–160. DOI: 10.1016/bs.aivir.2015.03.002.

[29] Simon Tavaré. „Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences“. In: *Lectures on Mathematics in the Life Sciences*. Bd. 17. American Mathematical Society, 1986, S. 57–86.

[30] Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding und Martin Steinegger. „Fast and Accurate Protein Structure Search with Foldseek“. In: *Nature Biotechnology* 42.2 (Feb. 2024), S. 243–246. ISSN: 1546-1696. DOI: 10.1038/s41587-023-01773-0. (Besucht am 01. 12. 2024).

[31] Gabriela Vieyres und Thomas Pietschmann. „Entry and Replication of Recombinant Hepatitis C Viruses in Cell Culture“. In: *Methods (San Diego, Calif.)* 59.2 (2013), S. 233–248. DOI: 10.1016/j.ymeth.2012.09.005.

[32] Scott C. Weaver und Nikos Vasilakis. „Molecular Evolution of Dengue Viruses: Contributions of Phylogenetics to Understanding the History and Epidemiology of the Preeminent Arboviral Disease“. In: *Infection, Genetics and Evolution* 9.4 (2009), S. 523–540. DOI: 10.1016/j.meegid.2009.02.003.

[33] Ziheng Yang. „Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods“. In: *Journal of Molecular Evolution* 39.3 (Sep. 1994), S. 306–314. ISSN: 1432-1432. DOI: 10.1007/BF00160154. (Besucht am 20. 12. 2024).

B Mathematische Formeln und Anwendung

B.1 Multiple Sequence Alignment mit MAFFT

MAFFT wird zur Erstellung von Multiplen Sequenzalignments (MSA) genutzt. Durch eine Fourier-Transformation wird eine effiziente Berechnung dabei ermöglicht. Die Qualität eines Alignments wird durch eine Punktzahl $\mathcal{S}$ bewertet, die die Übereinstimmung zwischen den Sequenzen quantifiziert:

$$\mathcal{S} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{ij} \cdot \text{Score}(i,j) \tag{B.1}$$

Hierbei ist w_{ij} ein Gewichtungsfaktor, der die Relevanz des Vergleichs zwischen Sequenz i und Sequenz j angibt, und $\text{Score}(i,j)$ ist die Ähnlichkeit zwischen den beiden Sequenzen basierend auf Substitutionsmatrizen wie BLOSUM oder PAM.

Transformation in das Frequenzspektrum MAFFT nutzt die Fourier-Transformation zur effizienten Berechnung der Ähnlichkeiten zwischen Sequenzen. Die Transformation ist definiert als:

$$\mathcal{F}(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-i\frac{2\pi kn}{N}} \tag{B.2}$$

Hierbei ist $\mathcal{F}(k)$ das Frequenzspektrum der Sequenz, $x(n)$ die diskrete Funktion der Sequenz an Position n, N die Länge der Sequenz und i die imaginäre Einheit. Die resultierenden Distanzwerte werden in einer Distanzmatrix $\mathcal{D}(i,j)$ zusammengefasst:

$$\mathcal{D}(i,j) = 1 - \frac{\sum_{k=1}^{L} \delta(x_k^i, x_k^j)}{L} \tag{B.3}$$

wobei $\delta(x_k^i, x_k^j)$ 1 ist, wenn die Aminosäuren an Position k in Sequenz i und j identisch sind, sonst 0. L ist die Länge des Alignments.

Praktische Anwendung: MAFFT kann auf der Kommandozeile folgendermaßen ausgeführt werden:

```
mafft [arguments] input > output
```

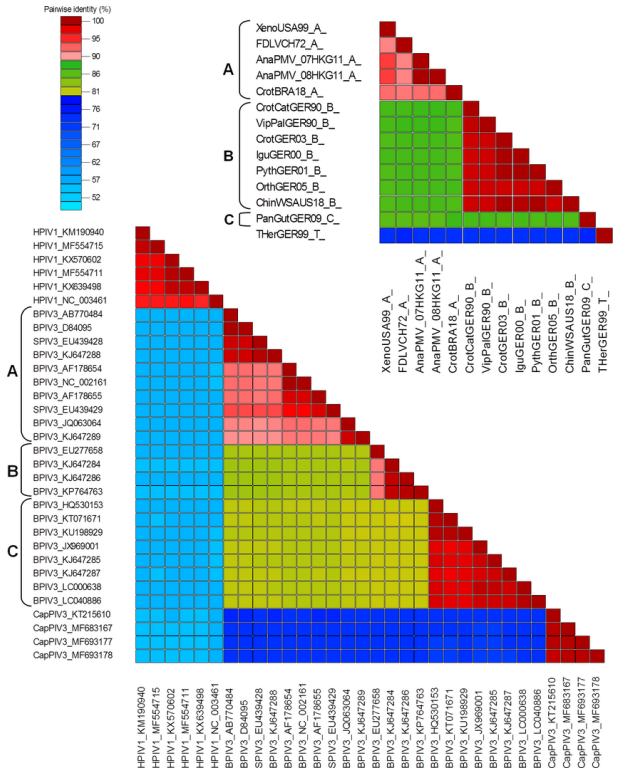

Abbildung B.1: Beispiel einer Distanzmatrix nach der Anwendung von MAFFT. [22]

B.2 Maximum-Likelihood-Methode und Substitutionsmodell

Die Maximum-Likelihood-Methode zielt darauf ab, die Baumtopologie T zu finden, die die Wahrscheinlichkeit der beobachteten Daten D maximiert, gegeben ein Modell M:

$$\mathcal{L}(M) = P(D \mid M) = \prod_{i=1}^{n} P(d_i \mid M) \tag{B.4}$$

Hierbei ist $\mathcal{L}(M)$ die Likelihood des Modells, und $P(d_i \mid M)$ die Wahrscheinlichkeit der Daten an Position i, gegeben das Modell M.

Substitutionsmodell GTR+I+G Das GTR+I+G-Modell berücksichtigt:

- *Unterschiedliche Substitutionsraten* zwischen Nukleotiden.
- *Invariante Positionen* (I), die konserviert bleiben.
- *Rate-Heterogenität* (G), die durch eine Gamma-Verteilung modelliert wird.

Die Gamma-Verteilung Γ beschreibt die Variabilität der Mutationsrate an verschiedenen Positionen:

$$\Gamma(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \tag{B.5}$$

wobei α und β die Form- und Skalierungsparameter sind.

Praktische Anwendung: Die Baumrekonstruktion mit IQ-TREE kann durch folgenden Befehl durchgeführt werden:

```
iqtree -s aligned_sequences.fasta -m GTR+I+G -bb 1000
```

Dabei führt IQ-TREE 1.000 Bootstrapping-Wiederholungen durch.

B.3 Berechnung der Root-Mean-Square Deviation (RMSD)

Die Root-Mean-Square Deviation (RMSD) misst die durchschnittliche Distanz zwischen entsprechenden Atomen zweier Proteinstrukturen nach optimaler Überlagerung:

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \|\mathbf{p}_i - \mathbf{q}_i\|^2} \tag{B.6}$$

Dabei sind $\mathbf{p}_i$ und $\mathbf{q}_i$ die Koordinaten des i-ten Atoms in den beiden Strukturen, und N ist die Gesamtzahl der Atome.

Beispiel zur RMSD-Berechnung: Gegeben sind zwei Atome mit den Koordinaten $\mathbf{p}_i = (1, 2, 3)$ und $\mathbf{q}_i = (2, 3, 4)$:

$$\|\mathbf{p}_i - \mathbf{q}_i\| = \sqrt{(2-1)^2 + (3-2)^2 + (4-3)^2} = \sqrt{3}. \tag{B.7}$$

Die RMSD ergibt sich durch das Mittel der quadrierten Abweichungen über alle Atome.