RUHR-UNIVERSITÄT BOCHUM

RUB

RUHR-UNIVERSITÄT BOCHUM

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

Tobias Polley

Wissenschaftliche Arbeit – 17. 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

Erklärung

Ich erkläre mich damit einverstanden, dass meine Abschlussarbeit am Lehrstuhl Bioinformatik dauerhaft in elektronischer und gedruckter Form aufbewahrt wird und dass die Ergebnisse aus dieser Arbeit unter Einhaltung guter wissenschaftlicher Praxis in der Forschung weiter verwendet werden dürfen. Weiterhin erkläre ich mich damit einverstanden, dass die Abschlussarbeit auf der Webseite des Lehrstuhls bzw. des Betreuers veröffentlicht werden darf.

Date	Tobias Polley

Inhaltsverzeichnis

Bemerkungen zum Entwurf der Arbeit

Die vorliegende Arbeit umfasst aktuell noch keine Grafiken, welche die Anschaulichkeit der Erläuterungen noch verbessern. Zudem ist aktuell noch der Umfang der Arbeit zu groß. Ich weiß leider nicht, an welchen Stellen ich die Arbeit noch kürzen könnte. Ich bitte um Feedback und Anregungen, um die Arbeit zu verbessern.

1 Einleitung

1.1 Vorstellung der Flaviviridae-Virusfamilie

Die Familie der Flaviviridae umfasst eine große und vielfältige Gruppe von einzelsträngigen RNA-Viren positiver Polarität, die sowohl Menschen als auch eine Vielzahl von Tieren infizieren können [29]. Zu den klinisch und ökologisch bedeutsamsten Vertretern gehören Erreger schwerwiegender Krankheiten wie Dengue-Fieber, Gelbfieber, Zika-Virus-Erkrankung und Hepatitis C [18]. Die Klassifikation der Flaviviridae unterteilt sich in vier Hauptgattungen: Flavivirus, Pestivirus, Pegivirus und Hepacivirus [29]. Neuere Entdeckungen, wie die Jingmenviren und großen Genom-Flaviviren (Large Genomes Flaviviruses, LGFs), haben die evolutionäre und ökologische Diversität dieser Familie erheblich erweitert [28].

Eine der zentralen Eigenschaften der Flaviviridae ist die Nutzung von Glycoproteinen zur Vermittlung des Viruseintritts in Zielzellen [22]. Diese Proteine sind essenziell für die Bestimmung der Wirtsspezifität, des Gewebetropismus und der Pathogenese [25]. Trotz ihrer Bedeutung bestehen noch erhebliche Lücken im Verständnis ihrer Struktur und Funktion, insbesondere bei nicht-klassifizierten Mitgliedern der Flaviviridae [19].

1.2 Bedeutung der Glycoprotein-Struktur für Evolution und Pathogenese

Die Glycoproteine der Flaviviridae sind hochinteressant in der Virusbiologie, z.B. sind sie Schlüsselstrukturen für die Interaktion mit dem Immunsystem des Wirts [9]. Sie umfassen hauptsächlich Klasse-II-Fusionsproteine, die für die Membranfusion verantwortlich sind, wie das E-Glycoprotein der Flaviviren [14]. Darüber hinaus existieren potenziell neue und unbekannte Mechanismen, wie sie bei Hepaciviren und Pegiviren im E1/E2-Komplex gefunden wurden [16]. Unterschiede in diesen Glycoproteinen spiegeln evolutionäre Anpassungen wider, die durch Mutationen, Rekombinationen und horizontalen Gentransfer vorangetrieben wurden [34].

Durch die Untersuchung der Glycoprotein-Divergenz können wichtige Einblicke in die Mechanismen gewonnen werden, die die ökologische Nischenanpassung und das Pathogenitätsprofil der Flaviviridae formen [19]. Insbesondere die Identifikation struktureller Merkmale, wie der hydrophoben Fusion-Loop-Region oder der transmembraner Domänen, liefert Hinweise auf die evolutionären und funktionellen Treiber dieser Proteine [21].

1.3 Zielsetzung und Vorgehensweise

Diese Arbeit hat zum Ziel, durch eine umfassende Analyse der Arbeit von *Mifsud, J.C.O., Lytras, S., Oliver, M.R. et al.* und dessen Methoden die Erkenntnisse über die evolutionäre Geschichte und Pathogenese der Flaviviridae gewinnen. Basierend auf dem Ansatz der referenzierter Arbeit werden phylogenetische Analysen, Proteinstrukturvorhersagen mittels ColabFold und ESMFold sowie Homologiesuchen mit Foldseek analysiert.

Die Arbeit ist wie folgt gegliedert: Zunächst werden in Kapitel 2 die eingesetzten bioinformatischen Verfahren und Werkzeuge beschrieben. In Kapitel 3 werden die wichtigsten Erkenntnisse zu phylogenetischen Beziehungen und Proteinstrukturen dargestellt. Kapitel 4 interpretiert die Ergebnisse hinsichtlich ihrer Implikationen für die Evolution und Pathogenese der Flaviviridae und diskutiert methodische Limitierungen sowie Perspektiven für zukünftige Forschung. Abschließend fasst Kapitel 5 die zentralen Ergebnisse zusammen und gibt einen Ausblick auf mögliche Anwendungen und zukünftige Fragestellungen.

2 Methoden

2.1 Phylogenetische Analyse

Für die phylogenetische Untersuchung der Flaviviridae-Familie wurden 458 vollständige Genomsequenzen aus öffentlichen Datenbanken wie GenBank extrahiert und sorgfältig kuratiert, um eine hohe Datenqualität sicherzustellen [19]. Als phylogenetischer Marker diente das Nichtstrukturelles Protein 5 (NS5)-Gen, welches für die RNA-abhängige RNA-Polymerase (RNA-abhängige RNA-Polymerase (RdRp)) kodiert und aufgrund seiner hohen Konservierung innerhalb der Flaviviridae ideal für phylogenetische Analysen ist [13].

Die Sequenzen wurden mit dem Programm Multiple Alignment using Fast Fourier Transform (MAFFT) zu einem Multiplen Sequenzalignment (Multiple Sequence Alignment (MSA)) ausgerichtet [11]. MAFFT ermöglicht durch effiziente Algorithmen und die Nutzung von Fourier-Transformationen eine genaue Ausrichtung großer Datensätze, wodurch konservierte und variable Regionen innerhalb der Sequenzen identifiziert werden können. Die detaillierten mathematischen Methoden zur Erstellung der MSAs sind im Anhang B beschrieben.

Anschließend wurde das MSA verwendet, um einen phylogenetischen Baum zu rekonstruieren. Hierfür kam die Maximum-Likelihood-Methode zum Einsatz, implementiert im Programm Efficient and Effective Phylogenetic Software Tool (IQ-TREE) [23]. Das General Time Reversible Model with Invariant Sites and Gamma Distribution (GTR+I+G)-Substitutionsmodell wurde gewählt, da es eine flexible Modellierung von Nukleotidsubstitutionen ermöglicht und für komplexe phylogenetische Analysen geeignet ist [31]. Die mathematischen Grundlagen der Maximum-Likelihood-Methode und des verwendeten Substitutionsmodells sind im Anhang B dargestellt.

Die Robustheit der phylogenetischen Baumtopologie wurde durch Bootstrapping mit 1.000 Wiederholungen getestet [6]. Dieses statistische Verfahren prüft die Zuverlässigkeit der Knoten im Baum, indem es wiederholt Stichproben aus den Daten zieht und die Baumrekonstruktion durchführt. Hohe Bootstrap-Werte (in der Regel über 70%) deuten auf eine starke Unterstützung der entsprechenden Knoten hin.

Durch diese methodische Vorgehensweise konnten die phylogenetischen Beziehungen innerhalb der Flaviviridae-Familie detailliert untersucht und Hauptkladen identifiziert werden, die mit den beobachteten Unterschieden in den Glycoprotein-Strukturen korrelieren.

2.2 Proteinstrukturvorhersage: ColabFold und ESMFold Methoden

Für die Vorhersage der dreidimensionalen Strukturen der Glycoproteine der Flaviviridae wurden zwei fortschrittliche bioinformatische Methoden eingesetzt: Collaborative Protein Folding (ColabFold) und Evolutionary Scale Modeling for Protein Folding (ESMFold). Beide basieren auf tiefen neuronalen Netzwerken, nutzen jedoch unterschiedliche Ansätze zur Proteinstrukturvorhersage und ergänzen sich somit in ihrer Anwendung.

ColabFold ist eine optimierte und zugängliche Implementierung von AlphaFold2, die es ermöglicht, Proteinstrukturvorhersagen effizient und ressourcenschonend durchzuführen [20]. AlphaFold2 hat das Feld der Proteinstrukturvorhersage revolutioniert, indem es tiefe neuronale Netzwerke mit evolutionären Informationen aus MSAs kombiniert, um hochpräzise Strukturvorhersagen zu generieren [10].

Der Vorhersageprozess mit ColabFold umfasst mehrere Schritte. Zunächst wird für jede Glycoprotein-Sequenz ein MSA erstellt, indem homologe Sequenzen aus großen Datenbanken wie UniProt Reference Clusters at 90% Identity (UniRef90), MGnify und der Big Fantastic Database (BFD) identifiziert werden. Dieses MSA dient dazu, evolutionäre Informationen zu extrahieren, indem es konservierte Positionen und ko-evolutionäre Signale zwischen Aminosäuren erkennt. Anschließend werden die Sequenz und das MSA in das AlphaFold2-Modell eingegeben. Das Modell besteht aus einem Evolutionary Transformer Module (Evoformer)-Modul, das Transformer-Architekturen verwendet, um Sequenzinformationen und MSA-Daten zu verarbeiten und langreichweitige Wechselwirkungen zwischen Aminosäuren zu modellieren. Schließlich sagt ein Strukturvorhersagekopf die dreidimensionalen Koordinaten der Proteinatome voraus, wobei sowohl geometrische als auch physikalische Constraints berücksichtigt werden. Die Qualität der Vorhersagen wird durch den Predicted Local Distance Difference Test (pLDDT) bewertet, der einen Vertrauenswert für jede Position im Protein liefert.

Im Gegensatz dazu verwendet ESMFold einen proteinsprachbasierten Ansatz, der direkt aus Einzelsequenzen lernt, ohne auf MSAs angewiesen zu sein [17]. ESMFold basiert auf dem ESM-2-Modell, einem großen Transformer-Sprachmodell, das auf Millionen von Proteinsequenzen trainiert wurde. Es nutzt Techniken aus der natürlichen Sprachverarbeitung (Natural Language Processing (NLP)), um Muster und Regularitäten in Proteinsequenzen zu erkennen. Die Proteinsequenz wird in das Modell eingegeben, das eine kontextabhängige Repräsentation jeder Aminosäure erzeugt. Diese Repräsentationen werden dann verwendet, um die dreidimensionalen Koordinaten der Proteinstruktur vorherzusagen, indem Abstände und Orientierungen zwischen Aminosäuren geschätzt werden. Obwohl ESMFold keine MSAs verwendet, kann es dennoch genaue Vorhersagen liefern, insbesondere bei Proteinen mit wenigen oder keinen homologen Sequenzen.

Beide Methoden wurden angewandt, um die Strukturen der Glycoproteine der verschiedenen Flaviviridae-Gattungen vorherzusagen. Durch den Einsatz von ColabFold konnten wir von den evolutionären Informationen profitieren, die in den MSAs enthalten sind, was insbesondere bei konservierten Proteinen zu präzisen Vorhersagen führt. ESMFold ergänzte diese Vorhersagen, indem es auch für hochdivergente Sequenzen zuverlässige Ergebnisse lieferte, bei denen wenige homologe Sequenzen verfügbar sind.

Die vorhergesagten Strukturen wurden anschließend validiert und verglichen. Eine Strukturüberlagerung der Modelle aus ColabFold und ESMFold ermöglichte die Beurteilung der räumlichen Übereinstimmung, und die Berechnung der Root-Mean-Square Deviation (Root-Mean-Square Deviation (RMSD)) lieferte quantitative Maße für Unterschiede zwischen den Strukturen. Die pLDDT-Werte wurden analysiert, um Bereiche hoher und niedriger Vorhersagegenauigkeit zu identifizieren. Besondere Aufmerksamkeit wurde konservierten strukturellen Motiven wie den hydrophoben Fusion-Loops geschenkt, die für die Funktion der Glycoproteine essenziell sind [21].

Die Vorhersagen der Glycoprotein-Strukturen ermöglichten es, funktionelle Domänen zu identifizieren und Unterschiede zwischen den Gattungen der Flaviviridae zu analysieren. So konnten wir beispielsweise Variationen in den Oberflächenexpositionen potenzieller Rezeptorbindungsstellen feststellen, was auf unterschiedliche Mechanismen des Viruseintritts und der Wirtsspezifität hindeutet [16].

Es ist jedoch zu beachten, dass die Genauigkeit der Vorhersagen von mehreren Faktoren abhängt. Bei Proteinen mit hoher Sequenzdivergenz oder wenigen verfügbaren homologen Sequenzen kann die Vorhersagegenauigkeit beeinträchtigt sein. Während ColabFold auf die Verfügbarkeit umfangreicher Sequenzdaten angewiesen ist, zeigt ESMFold Vorteile bei der Vorhersage von Proteinen mit geringer Homologie zu bekannten Strukturen. Dennoch ist die Interpretation der Vertrauenswerte und Vorhersagen im Kontext biologischer Funktion und experimenteller Validierung essenziell.

Die Kombination von ColabFold und ESMFold ermöglichte es, ein umfassendes Bild der Glycoprotein-Strukturen der Flaviviridae zu erhalten. Diese Strukturen bilden die Grundlage für weiterführende Studien zur Funktion, Evolution und potenziellen therapeutischen Zielstrukturen innerhalb dieser bedeutenden Virusfamilie.

2.3 Homologiesuche: Anwendung von Foldseek

Zur Untersuchung der evolutionären Beziehungen zwischen den Glycoproteinen der Flaviviridae und zur Identifizierung struktureller Homologien wurde das Programm Fast Protein Structure Search Tool (Foldseek) eingesetzt [32]. Foldseek ermöglicht einen schnellen und effizienten Vergleich von Proteinstrukturen auf Basis geometrischer und physikochemischer Eigenschaften, was besonders für große Datensätze und hochdivergente Sequenzen geeignet ist.

Die vorhergesagten Glycoprotein-Strukturen aus ColabFold und ESMFold wurden mit Foldseek gegen die Protein Data Bank (PDB) durchsucht, um potenzielle strukturelle Homologien zu bekannten Proteinen zu identifizieren. Dabei werden die dreidimensionalen Koordinaten der Proteine in vereinfachte Merkmalsrepräsentationen umgewandelt, die charakteristische strukturelle Merkmale erfassen [32]. Durch effiziente Algorithmen können so strukturelle Ähnlichkeiten auch bei geringer Sequenzidentität erkannt werden, was traditionelle sequenzbasierte Methoden wie BLAST nicht leisten [1].

Die Ergebnisse wurden anhand von Alignment-Scores und RMSD-Werten bewertet, um die Qualität der Übereinstimmungen zu quantifizieren. Hohe Übereinstimmungen deuteten auf konservierte Faltungsmuster hin, insbesondere die Klasse-II-Fusionsproteinfaltung, die für die Funktion der Glycoproteine essenziell ist [12]. Trotz hoher Sequenzdivergenz konnten gemeinsame strukturelle Merkmale identifiziert werden, was auf eine konservierte Funktion und einen gemeinsamen evolutionären Ursprung hindeutet [4].

Die Übereinstimmung zwischen zwei Strukturen wird durch die RMSD bewertet, wobei die detaillierte mathematische Herleitung im Anhang B erläutert wird.

Die Integration der Foldseek-Ergebnisse mit den phylogenetischen Analysen und den Proteinstrukturvorhersagen ermöglichte ein umfassendes Verständnis der evolutionären Beziehungen innerhalb der Flaviviridae. Diskrepanzen zwischen sequenzbasierten Phylogenien und strukturbasierten Homologien wurden analysiert, um mögliche Fälle von konvergenter Evolution oder funktioneller Anpassung zu identifizieren [5].

Durch diese methodische Vorgehensweise konnten funktionelle Domänen innerhalb der Glycoproteine identifiziert und neue Erkenntnisse über ihre evolutionäre Diversität gewonnen werden. Die strukturelle Homologiesuche mit Foldseek erwies sich somit als wertvolles Werkzeug zur Erweiterung des Verständnisses der Flaviviridae-Proteinstrukturen über die Grenzen sequenzbasierter Analysen hinaus.

3 Ergebnisse

3.1 Phylogenetische Strukturen und Klassifizierung

Die phylogenetische Analyse der Flaviviridae-Familie ergab eine Aufteilung in drei Hauptkladen, basierend auf den NS5-Sequenzen der 458 analysierten Virengenome. Durch die Verwendung des Maximum-Likelihood-Ansatzes mit dem GTR+I+G-Substitutionsmodell konnte ein robustes phylogenetisches Baumdiagramm erstellt werden, das durch hohe Bootstrap-Werte (90%) unterstützt wird [19].

Die erste Hauptklade umfasst die Gattung Flavivirus, zu der klassische Vertreter wie das Dengue-Virus, Zika-Virus und Gelbfieber-Virus gehören. Diese Gruppe zeigt eine hohe Sequenzkonservierung im NS5-Gen, was auf eine enge evolutionäre Verwandtschaft hindeutet [15]. Die Glycoprotein-Strukturen dieser Viren sind ebenfalls stark konserviert, insbesondere das E-Glycoprotein, das für die Membranfusion und den Viruseintritt in Wirtszellen essenziell ist [25].

Die zweite Hauptklade umfasst die Gattungen Pegivirus und Hepacivirus. Diese Gruppe zeigt eine größere genetische Diversität, insbesondere in den Glycoprotein-Genen. Die E1/E2-Glycoproteine dieser Viren weisen strukturelle Unterschiede zu den E-Glycoproteinen der Flaviviren auf, was auf unterschiedliche Mechanismen des Viruseintritts und der Wirtsspezifität hindeutet [33].

Die dritte Hauptklade vereint die Gattungen Pestivirus, Jingmenvirus und die großen Genom-Flaviviren (LGFs). Diese Gruppe zeichnet sich durch hochdivergente Sequenzen und eine bemerkenswerte strukturelle Vielfalt in den Glycoproteinen aus [27]. Die Pestiviren, die hauptsächlich Nutztiere infizieren, zeigen Unterschiede in den transmembranen Domänen und möglichen Rezeptorbindungsstellen, was auf spezifische Anpassungen an ihre Wirte hindeutet [30].

Die phylogenetische Topologie korreliert eng mit den beobachteten Unterschieden in den Glycoprotein-Strukturen, was darauf hindeutet, dass die evolutionäre Diversifikation dieser Proteine eine treibende Kraft in der Anpassung und Spezialisierung der Flaviviridae-Familie ist [19].

3.2 Glycoprotein-Divergenz: Unterschiede zwischen Gattungen

Die Analyse der Glycoproteine mittels ColabFold und ESMFold offenbarte sowohl konservierte als auch variable strukturelle Merkmale zwischen den verschiedenen Gattungen.

Die Flaviviren zeigen eine hohe Konservierung des E-Glycoproteins, insbesondere in der hydrophoben Fusion-Loop-Region und den transmembranen Domänen [25]. Die vorhergesagten Strukturen bestätigten die typische Klasse-II-Faltungsarchitektur mit drei Domänen, die hauptsächlich aus β-Faltblättern bestehen. Die Fusion-Loop, eine kritische Region für die Membranfusion, ist durch konservierte hydrophobe Aminosäuren charakterisiert, was ihre essenzielle Rolle unterstreicht [21].

Die Hepaciviren und Pegiviren zeigen eine einzigartige Organisation ihrer Glycoproteine, bestehend aus den E1- und E2-Proteinen, die Heterodimere bilden [33]. Die Strukturvorhersagen deuten auf signifikante Unterschiede in der räumlichen Anordnung hin, insbesondere in den Oberflächenexpositionen potenzieller Rezeptorbindungsstellen. Diese Unterschiede könnten die spezifische Wirtsspezifität und das breite Wirtspektrum dieser Viren erklären [16].

Die Pestiviren und Jingmenviren weisen die größte strukturelle Diversität auf. Ihre Glycoproteine zeigen Variabilität in der Anzahl und Position von transmembranen Domänen sowie in der Länge der extrazellulären Regionen [30]. Die Vorhersagen deuten darauf hin, dass diese Unterschiede funktionelle Anpassungen an spezifische Wirtsorganismen widerspiegeln könnten, möglicherweise durch die Interaktion mit unterschiedlichen Zellrezeptoren oder die Umgehung des Wirtsimmunsystems [8].

Trotz dieser Unterschiede wurden konservierte Motive identifiziert, insbesondere in den hydrophoben Kernen der Fusion-Loops und in bestimmten transmembranen Segmenten. Diese konservierten Regionen könnten unter starkem evolutionärem Selektionsdruck stehen, da sie für die grundlegenden Funktionen der Viren essenziell sind [21].

3.3 Spezifische Strukturmerkmale (z. B. Fusion-Loop und transmembrane Regionen)

Die detaillierte Untersuchung der Fusion-Loop-Regionen ergab, dass diese Bereiche aus konservierten hydrophoben Aminosäureresten bestehen, die für die Interaktion mit der Wirtszellmembran entscheidend sind [21]. In den Flaviviren ist die Fusion-Loop gut charakterisiert und spielt eine zentrale Rolle bei der Membranfusion unter sauren pH-Bedingungen im Endosom der Wirtszelle [25]. Die Strukturvorhersagen bestätigten die Anwesenheit der charakteristischen Schleifenstruktur, die von zwei konservierten Glycinresten flankiert wird, was für die Flexibilität und Funktion der Fusion-Loop wichtig ist [21].

Bei den Hepaciviren und Pegiviren zeigen die Fusion-Loop-Regionen Unterschiede in der Sequenz und möglichen Sekundärstruktur, was auf alternative Fusionsmechanismen hindeuten könnte [16]. Die E1/E2-Komplexe könnten einzigartige strukturelle Eigenschaften besitzen, die spezifisch für diese Gattungen sind und zur unterschiedlichen Pathogenität und Wirtsspezifität beitragen.

Die transmembranen Regionen der Glycoproteine wurden ebenfalls intensiv analysiert. Es wurde festgestellt, dass die Anzahl und Position der transmembranen Domänen zwischen den Gattungen variieren [30]. Flaviviren besitzen typischerweise eine einzelne transmembrane Domäne am C-Terminus des E-Glycoproteins, während Pestiviren und Hepaciviren mehrere transmembrane Segmente aufweisen können [24]. Diese Unterschiede könnten die Organisation der Glycoproteine in der viralen Membran beeinflussen und Auswirkungen auf die Virusassemblierung und Freisetzung haben [24].

Die Vorhersagen deuten auch auf die Präsenz von Signalpeptiden und Ankersequenzen hin, die für die richtige Lokalisierung und Funktion der Glycoproteine notwendig sind [2]. Variationen in diesen Bereichen könnten die Interaktion mit zellulären Faktoren beeinflussen und somit die Virusreplikation und Pathogenität modulieren.

Insgesamt liefern die Ergebnisse wichtige Einblicke in die strukturellen Merkmale der Glycoproteine der Flaviviridae und deren evolutionäre Anpassungen. Die Kombination aus phylogenetischer Analyse und Proteinstrukturvorhersage ermöglicht ein tieferes Verständnis der Mechanismen, die die Vielfalt und Spezialisierung dieser Virenfamilie vorantreiben [19].

4 Diskussion

4.1 Evolutionäre Bedeutung der Glycoprotein-Divergenz

Die Ergebnisse dieser Studie unterstreichen die zentrale Rolle der Glycoproteine in der Evolution und Anpassung der Flaviviridae-Familie. Die beobachtete Diversität der Glycoprotein-Strukturen spiegelt eine komplexe evolutionäre Dynamik wider, die durch Selektionsdruck, Wirtsspezifität und ökologische Nischen geprägt ist [19].

Die hohe Konservierung der hydrophoben Fusion-Loops und transmembranen Domänen über verschiedene Gattungen hinweg deutet darauf hin, dass diese Regionen essenziell für die grundlegenden Funktionen des Virus sind, insbesondere für die Membranfusion und den Eintritt in die Wirtszelle [25] [21]. Diese konservierten Strukturen stehen vermutlich unter starkem evolutionärem Selektionsdruck, da Änderungen in diesen Bereichen die Virulenz oder Infektiosität des Virus erheblich beeinträchtigen könnten.

Gleichzeitig weisen die variablen Regionen der Glycoproteine, wie beispielsweise die Oberflächenexpositionen und Glycosylierungsstellen, eine größere Diversität auf. Diese Unterschiede könnten auf adaptive Mechanismen zurückzuführen sein, die es den Viren ermöglichen, an verschiedene Wirtsorganismen anzupassen oder das Immunsystem des Wirts zu umgehen [16]. Beispielsweise könnten Variationen in den E1/E2-Komplexen der Hepaciviren dazu beitragen, die breite Wirtsspezifität und die Fähigkeit zur chronischen Infektion zu erklären [33].

Die strukturellen Unterschiede zwischen den Gattungen könnten auch das Ergebnis von Rekombinationsereignissen oder horizontalem Gentransfer sein, was zur Entstehung neuer Virusstämme mit unterschiedlichen Pathogenitätsprofilen führt [34]. Diese Mechanismen tragen zur genetischen Vielfalt der Flaviviridae bei und ermöglichen es den Viren, sich an veränderte Umweltbedingungen oder Wirtsimmunantworten anzupassen.

Insgesamt betonen die Ergebnisse die Bedeutung der Glycoprotein-Divergenz als treibende Kraft in der Evolution der Flaviviridae und liefern wichtige Einblicke in die molekularen Mechanismen der Virusadaption und Pathogenese.

4.2 Methodologische Limitierungen und Unsicherheiten

Obwohl die angewandten Methoden, insbesondere die Verwendung von ColabFold und ESMFold für die Proteinstrukturvorhersage, robuste Ergebnisse lieferten, gibt es einige methodologische Limitierungen, die berücksichtigt werden müssen.

Erstens hängt die Genauigkeit der Strukturvorhersagen von der Qualität der zugrunde liegenden Modelle und Algorithmen ab [10]. Insbesondere bei Proteinen mit geringer Sequenzähnlichkeit zu bekannten Strukturen kann die Zuverlässigkeit der Vorhersagen abnehmen. Dies betrifft vor allem die Glycoproteine der Pestiviren und Jingmenviren, bei denen wenig experimentelle Strukturdaten verfügbar sind.

Zweitens ist die Abhängigkeit von in silico Methoden ohne experimentelle Validierung eine potenzielle Quelle für Unsicherheiten. Obwohl die Vorhersagen durch konsistente Ergebnisse zwischen ColabFold und ESMFold gestützt werden, sind experimentelle Strukturanalysen, wie zum Beispiel durch Röntgenkristallographie oder Kryo-Elektronenmikroskopie, notwendig, um die Modelle zu bestätigen [3].

Drittens könnten die phylogenetischen Analysen durch Faktoren wie unvollständige Sequenzdaten, Rekombinationsereignisse oder unterschiedliche Evolutionsraten beeinflusst werden [6]. Die Verwendung eines einzelnen phylogenetischen Markers (NS5-Gen) könnte die Auflösung der phylogenetischen Beziehungen einschränken. Eine multimarkergestützte Analyse könnte hier zu detaillierteren Erkenntnissen führen.

4.3 Zukünftige Forschungsperspektiven

Die Ergebnisse dieser Studie eröffnen mehrere interessante Richtungen für zukünftige Forschung. Eine prioritäre Aufgabe ist die experimentelle Validierung der vorhergesagten Glycoprotein-Strukturen, insbesondere für weniger gut charakterisierte Gattungen wie die Pestiviren und Jingmenviren. Solche Studien könnten die Funktion spezifischer struktureller Merkmale bestätigen und neue Zielstrukturen für antivirale Therapien identifizieren.

Zudem könnten detaillierte Untersuchungen der variablen Regionen der Glycoproteine dazu beitragen, die Mechanismen der Wirtsspezifität und Immunabwehr besser zu verstehen. Dies ist besonders relevant für die Entwicklung von Impfstoffen und therapeutischen Antikörpern, die auf konservierte Epitope abzielen [7].

Die Erweiterung der phylogenetischen Analysen um zusätzliche genetische Marker und die Einbeziehung von Metagenomik-Daten könnten ein umfassenderes Bild der Evolution und Diversität der Flaviviridae liefern. Dies ist besonders wichtig angesichts der Entdeckung neuer Virusstämme und der potenziellen Risiken für die öffentliche Gesundheit.

Schließlich könnten die angewandten bioinformatischen Methoden weiter verbessert werden, um die Vorhersagegenauigkeit für Proteine mit geringer Sequenzähnlichkeit zu erhöhen. Die Integration von maschinellen Lernverfahren mit experimentellen Daten könnte hier neue Möglichkeiten eröffnen [26].

5 Schlussfolgerung

Die vorliegende Arbeit hat durch eine umfassende Analyse der Glycoprotein-Strukturen innerhalb der Flaviviridae-Familie neue Erkenntnisse über deren evolutionäre Geschichte und Pathogenese geliefert. Durch die Kombination von phylogenetischen Analysen basierend auf NS5-Sequenzen, Proteinstrukturvorhersagen mittels ColabFold und ESMFold sowie strukturellen Homologiesuchen mit Foldseek konnten sowohl konservierte als auch variable Merkmale der Glycoproteine identifiziert werden [19].

Die phylogenetische Analyse ergab eine Aufteilung der Flaviviridae in drei Hauptkladen, die eng mit den beobachteten Unterschieden in den Glycoprotein-Strukturen korrelieren. Konservierte Regionen wie die hydrophoben Fusion-Loops und transmembranen Domänen wurden über verschiedene Gattungen hinweg identifiziert, was auf ihre essenzielle Rolle in der viralen Funktion, insbesondere bei der Membranfusion und dem Eintritt in Wirtszellen, hindeutet [25] [21].

Gleichzeitig wurden signifikante strukturelle Variationen in den Glycoproteinen festgestellt, insbesondere in den E1/E2-Komplexen der Hepaciviren und Pegiviren sowie in den Glycoproteinen der Pestiviren und Jingmenviren. Diese Unterschiede könnten adaptive Mechanismen widerspiegeln, die es den Viren ermöglichen, sich an verschiedene Wirte anzupassen oder das Immunsystem zu umgehen [16].

Die Anwendung moderner bioinformatischer Methoden hat es ermöglicht, detaillierte Einblicke in die Struktur-Funktions-Beziehungen der Glycoproteine zu gewinnen und deren evolutionäre Diversifikation zu beleuchten. Trotz einiger methodologischer Limitierungen bieten die Ergebnisse eine solide Grundlage für zukünftige Forschungsarbeiten.

5.1 Zusammenfassung der Ergebnisse

Die Erkenntnisse dieser Arbeit haben mehrere praktische Implikationen. Die Identifizierung konservierter struktureller Merkmale wie der Fusion-Loops bietet potenzielle Zielstrukturen für die Entwicklung von breit wirksamen antiviralen Therapeutika und Impfstoffen [7]. Therapeutische Ansätze könnten darauf abzielen, diese konservierten Regionen zu blockieren und somit die virale Infektiosität zu reduzieren.

Die Variabilität in den Glycoprotein-Strukturen, insbesondere in den variablen Oberflächenregionen, eröffnet Möglichkeiten für die Entwicklung von gattungsspezifischen antiviralen Strategien. Ein besseres Verständnis der Mechanismen, die der Wirtsspezifität und Immunabwehr zugrunde liegen, könnte zur Entwicklung maßgeschneiderter Therapien beitragen.

Zukünftige Forschungsarbeiten sollten sich auf die experimentelle Validierung der vorhergesagten Strukturen konzentrieren, um die Genauigkeit der in silico Modelle zu bestätigen und deren funktionelle Relevanz zu untersuchen. Darüber hinaus wäre es sinnvoll, die phylogenetischen Analysen durch zusätzliche genetische Marker zu erweitern und Metagenomik-Daten einzubeziehen, um ein umfassenderes Bild der Evolution und Diversität der Flaviviridae zu erhalten.

Die Weiterentwicklung bioinformatischer Methoden, insbesondere in Bezug auf Proteinstrukturvorhersage und Homologiesuche, könnte die Genauigkeit und Zuverlässigkeit zukünftiger Analysen verbessern. Die Integration von maschinellem Lernen und experimentellen Daten bietet hier ein vielversprechendes Potenzial [26].

Insgesamt trägt diese Arbeit zum tieferen Verständnis der molekularen Mechanismen bei, die der Evolution und Pathogenese der Flaviviridae zugrunde liegen, und legt den Grundstein für zukünftige Forschungen, die letztlich zur Bekämpfung dieser bedeutenden Virenfamilie beitragen könnten.

Glossar

BFD Big Fantastic Database. 5

ColabFold Collaborative Protein Folding. 4–6

ESMFold Evolutionary Scale Modeling for Protein Folding. 4–6

Evoformer Evolutionary Transformer Module. 5

Foldseek Fast Protein Structure Search Tool. 6

GTR+I+G General Time Reversible Model with Invariant Sites and Gamma Distribution. 4, 21

IQ-TREE Efficient and Effective Phylogenetic Software Tool. 4

MAFFT Multiple Alignment using Fast Fourier Transform. 4

MSA Multiple Sequence Alignment. 4, 5, 20

NLP Natural Language Processing. 5

NS5 Nichtstrukturelles Protein 5. 4

PDB Protein Data Bank. 6

pLDDT Predicted Local Distance Difference Test. 5

RdRp RNA-abhängige RNA-Polymerase. 4

RMSD Root-Mean-Square Deviation. 5, 6, 21

UniRef90 UniProt Reference Clusters at 90% Identity. 5

Abbildungsverzeichnis

Tabellenverzeichnis

A List of AI-Tools

In the preparation of this thesis, several AI tools were utilized to enhance the research and writing process. I hereby declare that I have carefully checked all the suggestions generated by AI for correctness. The following tools were particularly instrumental:

ChatGPT Used for brainstorming research questions and to rephrase texts to improve readability. It was not used for crafting text bodies.

Grammarly Employed for proofreading and enhancing the overall writing quality.

DeepL Applied for translating key passages from English sources into German.

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] Cyrus Chothia und Arthur M. Lesk. „The Relation between the Divergence of Sequence and Structure in Proteins“. In: *EMBO Journal* 5.4 (1986), S. 823–826. DOI: 10.1002/j.1460-2075.1986.tb04288.x.

[5] W. Ford Doolittle. „Phylogenetic Classification and the Universal Tree“. In: *Science* 284.5423 (1999), S. 2124–2129. DOI: 10.1126/science.284.5423.2124.

[6] 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.

[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] Philippe Georgel, Catherine Schuster, Mirjam B. Zeisel, Françoise Stoll-Keller, Thomas Berg, Seiamak Bahram und Thomas F. Baumert. „Virus–Host Interactions in Hepatitis C Virus Infection: Implications for Molecular Pathogenesis and Antiviral Strategies“. In: *Trends in Molecular Medicine* 16.6 (Juni 2010), S. 277–286. ISSN: 1471-4914, 1471-499X. DOI: 10.1016/j.molmed.2010.04.003. (Besucht am 01. 12. 2024).

[9] 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.

[10] 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.

[11] 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.

[12] Margaret Kielian und Félix A. Rey. „Virus Membrane-Fusion Proteins: More than One Way to Make a Hairpin“. In: *Nature Reviews Microbiology* 4.1 (2006), S. 67–76. DOI: 10.1038/nrmicro1326.

[13] 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.

[14] 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.

[15] Goro Kuno und Gwong-Jen J. Chang. „Full-Length Sequencing and Genomic Characterization of Bagaza, Kedougou, and Zika Viruses“. In: *Archives of Virology* 152.4 (2007), S. 687–696. DOI: 10.1007/s00705-006-0903-z.

[16] 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.

[17] 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).

[18] 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.

[19] 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).

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] Zifang Shang, Hao Song, Yi Shi, Jianxun Qi und George F. Gao. „Crystal Structure of the Capsid Protein from Zika Virus“. In: *Journal of Molecular Biology* 430.7 (März 2018), S. 948–962. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2018.02.006. (Besucht am 01. 12. 2024).

[28] 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).

[29] 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.

[30] 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.

[31] 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.

[32] 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).

[33] 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.

[34] 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.

B Mathematische Formeln

B.1 Multiple Sequence Alignment mit MAFFT

MAFFT ist ein Tool zur Erstellung von Multiplen Sequenzalignments (MSA). 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.

Um Ähnlichkeiten zwischen Sequenzen effizient zu berechnen, transformiert MAFFT die Sequenzen mittels Fourier-Transformation in das Frequenzspektrum:

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

Dabei 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 Distanz zwischen Sequenzen wird durch eine Distanzmatrix $\mathcal{D}(i,j)$ berechnet:

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

Hier ist $\delta(x_k^i, x_k^j)$ eine Indikatorfunktion, die 1 ist, wenn die Aminosäuren an Position k in Sequenz i und j identisch sind, sonst 0. L ist die Länge der Ausrichtung.

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:

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

Dabei 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.

Das verwendete Substitutionsmodell GTR+I+G berücksichtigt unterschiedliche Substitutionsraten zwischen Nukleotiden, einen Anteil invarianter Positionen und gamma-verteilte Rate-Heterogenität. Die genaue Formulierung des Modells beinhaltet komplexe mathematische Gleichungen zur Beschreibung der Übergangswahrscheinlichkeiten zwischen Nukleotiden.

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

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

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \|\mathbf{p}_i - \mathbf{q}_i\|^2} \tag{B.5}$$

Hierbei sind $\mathbf{p}_i$ und $\mathbf{q}_i$ die Ortsvektoren des i-ten Atoms in den beiden zu vergleichenden Strukturen, und N ist die Gesamtzahl der Atome (oder C_α-Atome) im Vergleich. $\| \cdot \|$ bezeichnet die euklidische Norm.

Die RMSD-Berechnung ermöglicht die Quantifizierung der strukturellen Ähnlichkeit zwischen Proteinmodellen und dient als Kriterium für die Qualität der Strukturüberlagerung.