Koevolution in molekularen Komplexen

Zur Erlangung des akademischen Grades eines Doctor rerum naturalium (Dr. rer. nat.)
vom Fachbereich Biologie der Technischen Universität Darmstadt genehmigte Dissertation von
Dipl.-Biol. Philipp Weil aus Berlin

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Fachbereich Biologie
AG Computational Biology & Simulation

1. Referent: Prof. Dr. Kay Hamacher
2. Referent: Prof. Dr. Gerhard Thiel

Tag der Einreichung: 18.05.2012
Tag der Prüfung: 18.07.2012

Darmstadt 2012
D 17

Erklärung zur Dissertation

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Dissertation selbständig angefertigt habe. Sämtliche aus fremden Quellen direkt oder indirekt übernommene Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und noch nicht veröffentlicht.

Darmstadt, den 6. August 2012

_____

(Philipp Weil)

Die vorliegende Arbeit wurde im Zeitraum von April 2008 bis Juli 2012 unter der Leitung von Herrn Prof. Dr. Kay Hamacher am Institut für Mikrobiologie und Genetik des Fachbereichs Biologie der Technischen Universität Darmstadt in der AG Computational Biology and Simulation als Promotionsarbeit angefertigt.

Teile dieser Arbeit und Projekte, die während meiner Arbeit entstanden, sind bereits veröffentlicht worden.

P. Weil, F. Hoffgaard, K. Hamacher. *Estimating Sufficient Statistics in Co-Evolutionary Analysis by Mutual Information*, **Computational Biology and Chemistry** 33(6):440-444, 2009.

[1]P. Boba, P. Weil, F. Hoffgaard, K. Hamacher. *Co-Evolution in HIV Enzymes*, **Proc. of Bioinformatics 2010**, p. 39-47, A. Fred, J. Filipe, H. Gamboa (eds.)

[2]F. Hoffgaard, P. Weil, K. Hamacher. *BioPhysConnectoR: Connecting Sequence Information and Biophysical Models*, **BMC Bioinformatics**, 11:199, 2010.

[3]S. Bremm, T. von Landesberger, M. Hess, T. Schreck, P. Weil, K. Hamacher, *Interactive Visual Comparison of Multiple Trees*, **IEEE Conference on Visual Analytics Science and Technology (VAST2011)**, accepted

1 Zu dieser Arbeit konnte ich Expertise und den Code zur Berechnung der *mutual information* beitragen.

2 Zu dieser Arbeit konnte ich Teile des Programmcodes und der Dokumentation, vor allem für die Berechnung der *mutual inforamtion*, beitragen.

3 In dieser Arbeit konnte ich biologische Expertise und das Applikationsbeispiel beitragen.

Inhaltsverzeichnis

1 Zusammenfassung

Der Bereich der molekularen Evolution beschäftigt sich mit der Untersuchung der Änderung der Primärsequenz und den durch Sequenzänderung vermittelten Selektionsvorteil des molekularen Phänotyps. Dabei kann sich die molekulare (Ko)evolution zum Beispiel in kompensatorischen Mutationen manifestieren, die durch den Selektionsdruck an den beobachteten Positionen favorisiert wird. Mit dem Verständnis über Prozesse struktur- und dynamikrelevanter Mutationen können biochemische Interaktionen identifiziert werden, die evolutionär wichtig sind. Besondere Bedeutung gewinnen solche Erkenntnisse im Bereich der Resistenzentwicklung von Medikamenten, da die oben genannten molekularen Koevolutionsvorgänge Randbedingungen an die prinzipielle Evolvierbarkeit von Resistenzen stellen. So konnte zum Beispiel ein intramolekulares Cluster innerhalb der HIV-1 Protease identifiziert werden, das sich bei der Behandlung der Patienten durch Proteaseinhibitoren etabliert hat.

In dieser Arbeit sollen verschiedene koevolutionäre Prozesse in unterschiedlichen Molekülen (Ribosom [Kapitel 3] und der HIV-1 Protease [Kapitel 4]) untersucht werden. In Verbindung mit der biophysikalischen Annotation des Ribosoms (Kapitel 5) soll eine differenziertere Analyse der Koevolution ermöglicht werden.

Während dieser Arbeit hat sich als eine der größten Herausforderungen die Analyse von koevolutionären Datenmengen ergeben. Wie in Kapitel 2 erarbeitet wird, dient die so genannte *mutual information* (MI) zur Quantifizierung der Koevolution. Die Berechnung der MI führt aber gleichzeitig zu Datenvolumina von 10^4 und mehr Größen, während gleichzeitig phylogenetische Effekte eine nicht minder große Herausforderung darstellen. Als ein viel versprechender Weg solche Probleme anzugehen werden seit einiger Zeit in der Informatik so genannte *Visual Analytics*-Techniken diskutiert und entwickelt. Diese Idee aufgreifend wird in Kapitel 6 eine entsprechende Software vorgestellt.

2 Das Nullmodell für die *Mutual Information* (MI)

Dieses Kapitel ist in [136] veröffentlicht worden.

2.1 Einleitung

In der Bioinformatik werden unter anderem evolutionäre Signale in Datensätzen von Sequenzen einzelner Biomoleküle (Protein- oder Nukleotidsequenzen) gesucht. Molekulare Koevolution entsteht aus der gemeinsamen Evolution mehrerer Residuen innerhalb eines Moleküls oder zwischen interagierenden Molekülen. Die stärksten koevolutionären Signale wurden in α-Helices identifiziert und dort den α-helikalen-*stacking* Interaktionen zugeordnet [24]. Andere Treiber für Koevolution sind sekundäre, tertiäre und quartäre Strukturen innerhalb des Moleküls, sowie Residuen die an Bindestellen und im aktiven Zentrum lokalisiert sind. Ein Ansatz diese Signale zu quantifizieren ist die *mutual information* (MI) (Kapitel 2.2.2), die bisher vor allem in der Signalverarbeitung schon erfolgreich eingesetzt wurde [67, 108, 3, 110].

Obwohl die MI oftmals genutzt wird, um koevolutionäre Positionen in Sequenzalignments zu identifizieren, gibt es nur wenige Studien, die sich mit dem *finite-size* Effekt von empirischen Daten beschäftigen. Diese Studien bestehen aus einigen theoretischen Modellen [91], die wenig Applikationsmöglichkeiten auf realen Daten aus biologischen Systemen bieten. Trotz alledem wurden Versuche unternommen das MI-Signal zu verbessern [56, 22, 53]. Die meisten dieser Versuche gewichten oder kalibrieren die berechneten Ergebnisse neu [7], um drei Effekten entgegenzuwirken:

1. *finite-size* Effekte, Effekte durch zu kleine Datensätze
2. phylogenetische Effekte, ein Überschätzen der MI, die zwischen den untersuchten und nah verwandten Spezies auftreten können
3. Der Grad der Konserviertheit einer Spalte in einem Multiplen-Sequenz-Alignment (MSA).

Obwohl immer größere Sequenzzahlen verfügbar sind, sind die Datensätze zum Teil dominiert vom *finite-size*-Effekten. Somit ergibt sich die Frage, wie viele Daten benötigt werden um signifikante Signale mit Hilfe der MI zu identifizieren. Mit der Antwort könnten Gewichtung und Kalibrierung umgangen werden [7].

2.2 Methoden

2.2.1 Alignments

Alignments aus biologischen Sequenzen entstehen durch die Ausrichtung der Sequenzen, sodass ähnliche Sequenzteile untereinander stehen. Diese Ausrichtung wird mit Hilfe von Algorithmen, die mit sogenannten Score-Matrizen eine Wertung der untereinander stehenden Residuen vornehmen. Dabei werden gleiche Residuen (*match*) positiv bewertet und ungleiche Residuen (*mismatches*) durch einen negativen Score bestraft. Daraus ergibt sich, dass über die Bestimmung der größte Score, die beste Übereinstimmung der untersuchten Sequenzen gefunden

werden kann. Die Score-Matrizen werden aus empirisch bestimmten Mutationswahrscheinlichkeiten abgeleitet, wie zum Beispiel der BLOSSUM62, die in den Alignment-Algorithmen implementiert sind. Bei dieser Bewertung der Sequenzen kann es dazu kommen, dass einzelne Teilstücke einer Sequenz nicht gepaart werden können und dadurch Lücken, so genannte Gaps, entstehen. Die Gap-Symbole in einem Alignment ergeben sich also durch die Unzuordbarkeit einzelner Positionen oder Positionsblöcken bei der Berechnung des Alignments. Genetisch wird dies durch Insertion und Deletion begründet.

Die Alignments der rRNA und der ribosomalen Proteine wurden jeweils mit Hilfe von `clustalw` [130, 28, 78] mit Standardparameter erstellt.

2.2.2 *Mutual Information*

Die MI ist ein Maß in der Informationstheorie, mit dessen Hilfe verallgemeinerte Korrelationen in Daten identifiziert werden können. Die MI leitet sich aus der Shannon-Entropie [123] ab, die definiert ist als:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{1}$$

dabei stellt x alle möglichen Realisierungen einer Zufallsvariablen X dar. Die Shannon-Entropie ist daher ein Maß, das etwas über den Informationsgehalt innerhalb einer beobachteten Sequenz von Realisierungen einer Zufallsvariable aussagt. Die Entropie wird nämlich genau dann maximal, wenn in einer Sequenz alle Realisierungen vorkommen und gleichverteilt vorliegen [85]. Die Information, die durch die Shannon-Entropie bestimmt wird, gibt Auskunft über den Grad der Variabilität innerhalb einer beobachteten Sequenz. Damit ergibt sich für die minimale Information eine Sequenz, die nur aus einer einzigen Realisierung besteht und somit keine Variabilität enthält. Wird dieser Fall beobachtet ergibt sich eine Shannon-Entropie von Null.
Um eine verallgemeinerte Korrelation zwischen zwei Datensätzen zu bestimmen, wird die MI verwendet, die wie folgt definiert ist:

$$\mathrm{MI} = H(X) + H(Y) - H(X, Y) \tag{2}$$

Eine alternative, aber äquivalente Definition der MI ist nach MacKay [85]:

$$\mathrm{MI}(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \cdot \log_2 \frac{p(x, y)}{p(x) p(y)} \tag{3}$$

dabei sind x und y auch hier die Realisierungen der Zufallsvariablen X und Y.

Betrachten wir die MI am Beispiel von biologischen Sequenzen, wie einem MSA von Proteinsequenzen, dann quantifiziert die MI die korrelierte Änderung zweier Spalten innerhalb des Alignments. In diesem Alignment stellen die Realisierungen x und y einzelne Symbole entweder des Standardaminosäurealphabets der 20 proteinogenen Aminosäuren (AS) aus in den Zufallsvariablen X und Y dar, oder des Basenalphabets in DNA oder RNA-Alignments. Dabei entsprechen die Zufallsvariablen X und Y jeweils einer Spalte des Alignments. Eventuell wird dieses Alphabet erweitert um ein Gap-Symbol (“-“). Um eine korrelierte Änderung mit Hilfe der

MI zu quantifizieren, betrachten wir zwei verschiedene Spalten des Alignments, also zwei unterschiedliche Positionen X und Y innerhalb des beobachteten Moleküls.
Da die MI die korrelierte Änderung der beiden betrachteten MSA-Spalten quantifiziert, muss die MI Null sein, wenn die Spalten nicht miteinander korrelieren. Diese Behauptung lässt sich durch die Definition von Unabhängigkeit in Zufallsexperimenten zeigen. Hier würde gelten: $p(x,y) = p(x) \cdot p(y)$, setzt man diesen Zusammenhang in Gleichung 3 ein, wird der logarithmische Term gleich eins und somit die MI immer null.
Für das theoretische Maximum der MI betrachten wir Gleichung 2. Hier wird der Informationsgehalt der beiden beobachteten MSA-Spalten zunächst addiert und anschließend wird der Informationsgehalt der gemeinsamen Entropie subtrahiert. Aus diesem Zusammenhang wird klar, das für die maximale MI die Einzelentropien $H(X)$ und $H(Y)$ maximiert und die gemeinsame Entropie $H(X,Y)$ minimiert werden müssen. Für die Maximierung der Shannon-Entropie müssen alle AS vorkommen und gleichverteilt vorliegen [85]. Ist das der Fall, dann folgt für das gemeinsame Auftreten von Realisierungen in der Berechnung von $H(X,Y)$, dass minimal so viele verschiedene gemeinsame Realisierungen auftreten, wie AS vorkommen und maximal alle AS mit allen AS gemeinsam auftreten. Im minimalen Fall wird $H(X,Y)$ dann genau so groß wie $H(X)$ und $H(Y)$, sodass gilt: $H(X,Y) = H(X) = H(Y)$. Im maximalen Fall existiert die quadratische Anzahl von Realisierungsmöglichkeiten N, sodass gilt: $H(X,Y) = H(X) + H(Y)$. Betrachtet man Gleichung 1 ergibt sich für die maximale MI, wenn die Realisierungsmöglichkeiten N_X der Zufallsvariablen X und N_Y der Zufallsvariablen Y gleich groß sind, folgender Zusammenhang:

$$max\,\mathrm{MI}(X,Y) = H(X) = \log_2 N_X \tag{4}$$

Sind die Realisierungsmöglichkeiten der N_x und N_y unterschiedlich groß ergibt sich folgender Zusammenhang für die maximale MI:

$$max\,\mathrm{MI}(X,Y) = \log_2 N_X + \log_2 N_Y - \log_2 min\,(N_X, N_Y), \tag{5}$$

Für den Vergleich innerhalb eines Protein-Alignments folgt für die MI daher ein Bereich zwischen Null und $\log_2 N_X$ ~4,32, da $\log_2 N_Y - \log_2 min\,(N_X, N_Y) = 0$ ist.

2.2.3 Das *Shuffle*-Nullmodell

Die Sequenzen für die Protein-Alignments wurden aus chromosomalen Datensätzen von 778 Mikroorganismen extrahiert. Dabei wurden die Datensätze der Genome aus der GenBank [14] verwendet. Diese Genomsequenzen wurden mit Hilfe von `Bio-Python` [32] in ihre Proteinsequenzen transkribiert und in allen möglichen Leserahmen übersetzt. In diesen transkribierten Sequenzen wurde dann mit Hilfe der Hidden Markov Modelle (HMM) (Kapitel 3.2.3) nach Signaturen von ribosomalen Protein-Domänen gesucht. Die Parameter der HMMs stammen aus der Pfam-Datenbank [48]. Aus den gefundenen Sequenzen wurden nur solche behalten, die einen Signifikanzwert (E-Value) von $\leq 10^{-5}$ aufwiesen, um sicherzustellen, dass so wenige falsch positive Sequenzen identifiziert wurden wie möglich. Die anschließenden Alignments wurden mit Hilfe von `clustalw` [28, 78] mit Standardparametern durchgeführt. Zu diesem Datensatz wurden Sequenzen des viralen Ionenkanals Kcv [128] und der HIV-1 Protease [61, 62] hinzugefügt.

Die MI wurde für alle Alignments nach Gleichung 3 berechnet. Eine gewichtete MI der Alignments wurde für die Sequenzen von gleichen Spezies durchgeführt, um der systematischen Verzerrung durch häufig sequenzierte Organismen entgegenzuwirken. Wenn zum Beispiel zehn Sequenzen des gleichen Organismus für ein Protein gefunden wurden, wurde jede nur mit einem Zehntel in der Frequenzbestimmung von $p(x)$, $p(y)$ und $p(x, y)$ berücksichtigt.

Um die eingangs erwähnten Effekte (Kapitel 2.1: *finite-size* Effekte, phylogenetischer Effekt und der Grad der Konserviertheit) zu unterdrücken, gibt es unterschiedliche Verfahren (Kapitel 3.2.2), diese Effekte durch Korrekturterme in der Berechnung der MI zu berücksichtigen. Weiterhin kann eine Korrektur auch empirisch durch computerbasierte Nullmodelle erreicht werden. In dieser Studie wurden zwei dieser Nullmodelle verglichen:

1. artifizielle Sequenzen aus zufällig generierten Buchstaben. Dabei bestand das Alphabet der Buchstaben aus der Einbuchstabenkodierung der 20 Standardaminosären. An jeder Position der Sequenz wurde der Buchstabe über eine gleichverteilte Wahrscheinlichkeit randomisiert generiert. Daraus ergibt sich eine Maximierung für $H(X)$ und $H(Y)$. Dieses Modell zeigt das statistische Rauschen des *finite-size* Effektes und ist analog zu dem Modell von Mahony et al. [87]. Um den Einfluss des Gap-Symbols zu zeigen, wurde zusätzlich ein 21ster Buchstabe (das Gap-Symbol "-") mit einer Wahrscheinlichkeit von $p(\text{"-"}) \sim 38\,\%$ eingeführt. Diese Wahrscheinlichkeit wurde aus den vorliegenden Alignments als typische Gap-Häufigkeit ermittelt.

2. randomisierte Spalten reeller Sequenzen, die aus den Alignments der untersuchten Biomoleküle generiert wurden. Dafür wurden die Alignmentspalten, wie von Hamacher [63] vorgeschlagen, durchmischt. Durch diese Methode bleiben die Häufigkeiten von $p(x)$ und $p(y)$ sowie alle vorhandenen phylogenetischen oder statistischen Eigenschaften erhalten. Die Abhängigkeit in $p(x, y)$ hingegen wird durch das Durchmischen aufgehoben und würde bei einem unendlich großen Datensatz zu einer MI von Null führen, da dann $p(x, y) = p(x) \cdot (y)$ exakt gelten würde. Daher ergibt sich aus diesem *Shuffle*-Nullmodell der *finite-size* Effekt, denn jede Abweichung der MI von Null in einem solchen unendlich großen Datensatz ist bedingt durch diesen Effekt. In diesem Modell wird implizit auch der Selektionsdruck jeder Position (ausgedrückt durch $H(X)$ und $H(Y)$) erhalten, der bei einem theoretischen Modell, wie dem von Martin et al. [91], zerstört würde.

Für die Berechnungen dieser Studie haben wir die Software `R` [113] in Kombination mit den Paketen `BioPhysConnectoR` [68] und `bio3d` [57] verwendet. Die Alignmentspalten wurden mit dem `R`-Befehl `sample()` durchmischt.

2.3 Resultate

Um den Einfluss des Hintergrundrauschens zu untersuchen und daraus die benötigte Anzahl an Sequenzen zu ermitteln, damit dieser Effekt hinreichend stark unterdrückt wird, werden die Einflüsse der beiden Nullmodelle verglichen. Dazu wird die Anzahl N von Sequenzen variiert, die zufällig erzeugt werden und ermitteln die durchschnittliche MI der *Shuffle*-Läufe. Im ersten Nullmodell werden dazu N zufällige AS-Paare erzeugt und die MI berechnet. Im zweiten Nullmodell werden die Spalten des Alignments mit Hilfe des `sample`-Algorithmus von `R` [113] durchmischt und die ersten N Sequenzen werden dann für die Berechnung der MI verwendet.

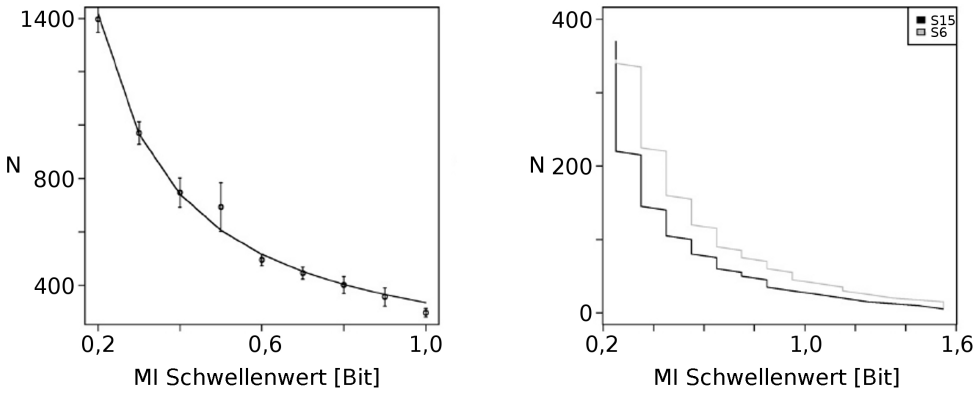

Abbildung 1: (A) Das erste Nullmodell mit Zufallssymbolen; die Anzahl der Sequenzen N, die benötigt werden, damit 50 % der berechneten MI-Werte unterhalb des Schwellenwertes auf der Abszisse liegen. Jeder Datenpunkt repräsentiert den Mittelwert aus 10 000 Iterationen. (B) Für die ribosomalen Proteine S6 und S15 zeigt der Plot die Anzahl an zufällig gezogenen Sequenzen N, die benötigt wurden, damit 95 % der berechneten MI-Werte unterhalb des auf der Abszisse gezeigten Schwellenwertes liegen.

Um die Abhängigkeit eines beliebig gewählten Schwellenwerts der durchschnittlichen MI des *Shuffle*-Nullmodells im Zusammenhang mit der benötigten Sequenzzahl zu illustrieren, ist in Abbildung 1 (A) die Sequenzzahl gegen den Schwellenwert der MI-Berechnung aufgetragen, die benötigt wurde, damit mindestens 50 % von 10 000 wiederholten Experimenten einen MI-Wert kleiner gleich dem Schwellenwert aufwiesen. In Abbildung 1 (B) ist das Ergebnis der Studie für das zweite Nullmodell gezeigt, hier sind allerdings diejenigen Sequenzzahlen gezeigt, bei denen 95 % unter dem Schwellenwert lagen, um die Unabhängigkeit der funktionellen Form zu zeigen. Beide Abbildungen zeigen, dass der Zusammenhang zwischen der Anzahl zufällig erzeugter Sequenzen und dem MI-Schwellenwert einem exponentiellen Verlauf folgt. Je niedriger die zufällig erzeugte MI als Hintergrundrauschen sein soll, desto höher muss demnach die Anzahl der verwendeten Sequenzen sein.

Die Abhängigkeit der berechneten MI aus den zufällig erzeugten Buchstabenpaaren und der Anzahl an generierten Paaren ist in Abbildung 2 (A) gezeigt. Hier sind die Mediane, der mit Hilfe des Nullmodells berechneten MI für 10 000 Iterationen über der Anzahl an verwendeten Buchstabenpaaren gezeigt. In Abbildung 2 (B) ist der doppelt logarithmische Plot für die Daten des Insets aus Abbildung 2 (A) gezeigt. Zu den Ergebnissen der ribosomalen Protein-

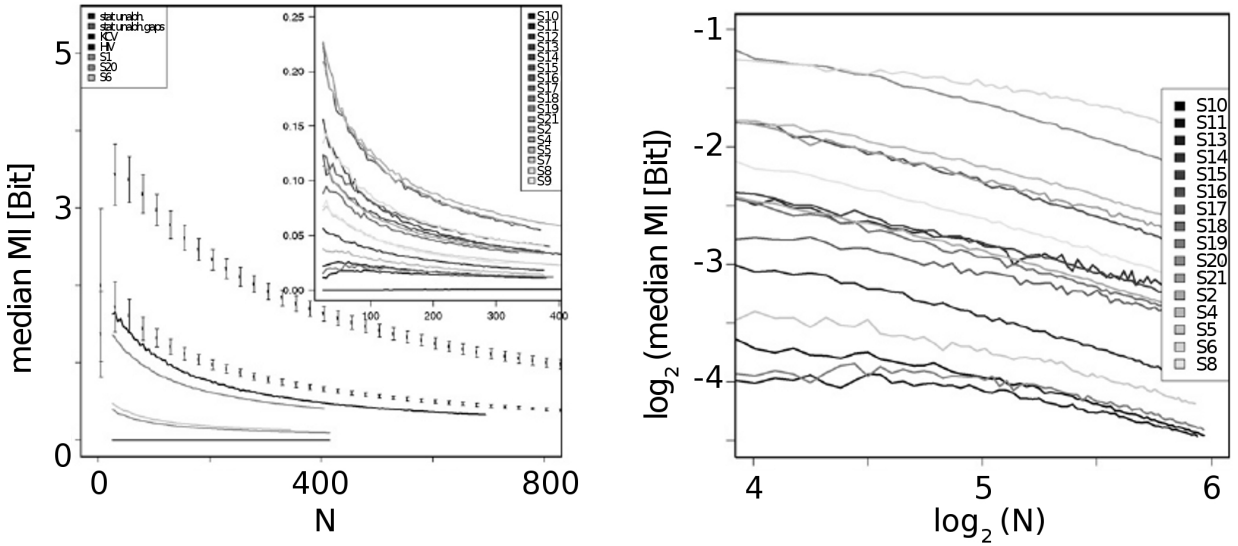

Abbildung 2: (A) Anzahl der Sequenzen N, die zur Berechnung des Medians der MI nach dem Randomisieren der Sequenzspalten des Original-Alignments, benutzt wurden. Die Datenpunkte mit Fehlerbalken zeigen den Verlauf der statistisch unabhängig erzeugten AS-Paare. Der Gap-Gehalt des Original-Alignments wurde benutzt um ein weiteres statistisch unabhängiges Modell zu berechnen, das Gaps als weiteres Symbol enthält (ohne Gaps (”stat. unabh.”) und mit einem Gap-Gehalt von $\sim$38 % (”stat. unabh. gaps”)). Für jedes N wurde die korrespondierende Anzahl an AS-Paare erzeugt und die MI wurde für jeweils 1 000 solcher Pseudo-Alignments berechnet. Der Median für jedes N wurde aufgetragen und mit Fehlerbalken der Standardfehler bei 1 000 Wiederholungen angezeigt. Inset: Wie im Hauptplot wurde hier die MI für jedes Alignment mit Hilfe des Shuffling-Nullmodells berechnet und der Median über der Anzahl N an, für die Berechnung verwendeten Sequenzen aufgetragen. (B) enthält den gleichen Plot wie das Inset aus (A) mit logarithmierte MI_{ij}-Werten (i, j bezeichnen hier die Spaltenindices des Alignments.)

Alignments wurden auch die Ergebnisse von Protein-Alignments für Kcv und HIV-1 Protease gezeigt. In Abbildung 2 (A) sind die mit Fehlerbalken eingezeichneten Ergebnisse die Verläufe der MI-Berechnungen für die zufällig generierten AS-Paare als obere Grenze gezeigt. Für einige Moleküle wurden nicht genügend Sequenzen gefunden, sodass einige Datensätze durch ein kleines N beschränkt sind. Trotz dieser Tatsache ist ein gleicher Verlauf der Untersuchung für alle Modelle zu erkennen. Aus den Abbildungen 1 und 2 lässt sich ein Zusammenhang erkennen zwischen der Anzahl der Sequenzen N und dem MI-Grenzwert t, der mit N unabhängig erzeugten Buchstaben zufällig erzeugt werden kann. Dieser Zusammenhang lässt sich wie folgt beschreiben:

$$N = N_0 t^{\alpha} \tag{6}$$

Aus diesem Zusammenhang wurde der Exponent α für die ribosomalen Proteine bestimmt. Dieser Exponent α zeigt im Mittel über alle Proteine einen Wert von -0,494±0,006, also im Wesentlichen einen Wurzelzusammenhang in Gleichung 6. Für den Vorfaktor N_0 wurden in den unterschiedlichen Proteinen verschiedene Werte gefunden, dieser Vorfaktor scheint aber nicht von der Größe der Proteine abhängig zu sein [Daten nicht gezeigt].

2.4 Diskussion

Die MI wurde als Maß der Koevolution zwischen verschiedenen Positionen in Alignments verschiedener Biomoleküle angewandt und zwei unterschiedliche Nullmodelle verwendet. Trotz phylogenetischer Beziehungen, der Überrepräsentation von gut studierten Organismen (wie z.B. *E. coli*) und einem variierenden Grad der Konserviertheit der AS innerhalb der verschiedenen Proteine haben wir ähnliche Fluktuationen in der MI auf Grund von *finite-size* Effekten feststellen können, wie es das Nullmodell statistisch unabhängig erzeugter AS-Paare zeigt.

Das hier vorgestellte *Shuffle*-Nullmodell kann also eine mittlere zufällige MI berechnen, die mit den zu Grunde liegenden Verteilungen aus realen Datensätzen entstehen kann und liefert so eine Abschätzung des *finite-size* Effektes unter Beibehaltung verschiedener anderer Effekte, wie zum Beispiel des phylogenetischen Effektes oder des lokalen Selektionsdrucks (repräsentiert durch $H(X)$ und $H(Y)$).

Die Untersuchung an weiteren Biomolekülen, wie dem viralen Kaliumkanal Kcv und der HIV-1 Protease, lässt den Schluss zu, dass die Ergebnisse für ein breites Spektrum an Proteinen zutreffen.

Die Abhängigkeit der MI in Bezug auf den *finite-size* Effekt ist in allen hier untersuchten Datensätzen so ähnlich, dass sich daraus für die Abschätzung der benötigten Größe des untersuchten Datensatzes folgendes Vorgehen ergibt:

- Anfertigen des *Shuffle*-Modells für ca. 100 Sequenzen.
- Berechnung des Medians der MI oder eines anderen Maßes, wie in Abbildung 1.
- Anwendung der Gleichung 6. Dies erlaubt eine Abschätzung der Anzahl an benötigten Sequenzen, um ein signifikantes MI-Signal zu erhalten.

- Alternativ kann die Anzahl an vorhandenen Sequenzen in Gleichung 6 eingesetzt und ein Schwellenwert ermittelt werden, der Auskunft darüber erteilt, ob ein Wert durch das Rauschen des kleinen Datensatzes erzeugt wurde.

Dieser Ansatz ähnelt einem Bootstrap-Verfahren [42, 142, 69, 75, 124]. In dem hier vorliegenden Fall „bootstrappen“ wir die Statistik des Hintergrunds, um die Relevanz des Signals zu zeigen.

Mit den gezeigten Ergebnissen muss der Hinweis von Gloor et al. [53], dass ein Datensatz aus 125 Sequenzen eine ausreichende Statistik für die Berechnung der MI bietet, als sehr optimistisch angesehen werden. Nach den vorliegenden Ergebnissen sind eher 200-300 Sequenzen nötig. Das hier vorgestellte *Shuffle*-Nullmodell wurde in das `R`-Paket `BioPhysConnectoR` [68] implementiert.

3 Ribosomale *Mutual Information*

3.1 Einleitung

Das Ribosom ist ein Molekülkomplex, der in allen lebenden Organismen die Transkription der messenger-Ribonukleinsäure (mRNA) in Proteine katalysiert. Das Ribosom besteht zum Großteil aus der ribosomalen RNA (rRNA), die als Peptidyltransferase die Verknüpfung der Aminosäuren (AS) katalysiert. Außerdem sind an der rRNA verschiedene ribosomale Proteine assoziiert. Das Ribosom unterscheidet sich in seinem Aufbau in Prokaryoten, Eukaryoten und Archaeen. Die Ribosomen wurde dabei historisch über die Dichtezentrifugation beschrieben, wodurch sich die Nomenklatur über die Swedbergeinheit etabliert hat. Die drei Domänen des Lebens haben unterschiedlich große Ribosomen, so findet sich in Prokaryoten ein 70 S Ribosom und in Eukaryoten ein 80 S Ribosom. Das in Archaeen gefundene Ribosom ähnelt strukturell dem eukaryotischen Ribosom ist aber in der Komposition aus den ribosomalen RNA molekülen 16 S, 23 S und 5 S rRNA dem baktereiellen Ribosom sehr ähnlich [82]. Das Ribosom besteht jeweils aus zwei Untereinheiten, einer großen und einer kleinen, die intrazellulär auch dissoziiert vorliegen können. In der Proteinbiosynthese werden die AS mit Hilfe von transfer-RNAs (tRNAs) in das Ribosom integriert. Dabei assoziiert die tRNA an der A-Site des Ribosoms und wird in die P-Site transportiert, in der die AS der tRNA kovalent an das naszierende Protein gebunden wird. Die leere tRNA wird schließlich an der E-Site des Ribosoms wieder freigesetzt. Die Peptidyltransferaseaktivität des Ribosoms befindet sich in der 50 S Untereinheit des bakteriellen Ribosoms. Eine Übersicht des strukturellen Aufbaus und der Funktion ist von Garrett et al. aufgestellt worden [52]. Die an die rRNA assoziierten Proteine sind durch die Untereinheit und ihre Größe definiert. So ist das L1 Protein in der großen Untereinheit (L für das englische *large*) zu finden und ist das größte Protein in der Analyse mit Hilfe von 2D-Gelelektrophorese. Das S21 Protein ist nach der allgemeinen Nomenklatur in der kleinen (S für das englische *small*) Untereinheit zu finden und zeigt sich an Position 21 in der 2D-Gelelektrophorese.

Die zentrale Rolle in der Proteinbiosynthese macht das Ribosom vor allem in der Bekämpfung von pathogenen Mikroorganismen zu einem primären Ziel für die medikamentöse Therapie, in der vor allem Antibiotika verwendet werden. Antibiotika können dabei in fast jedem Schritt des Zellzyklus in den Anabolismus der Zelle eingreifen, zum Beispiel durch Inhibition verschiedener Enzyme, die essentiell für das Wachstum oder das Überleben der Zelle sind [19, 39]. Viele Antibiotika inhibieren zum Beispiel die Proteinbiosynthese durch Bindung an das Ribosom. So entfalten etwa Tetracyclin und andere Antibiotika durch Bindung an die kleine ribosomale Untereinheit [17] ihre Wirkung. So genannte Makrolit-Antibiotika binden an der großen ribosomalen Untereinheit und verhindern das Austreten des synthetisierten Proteins [119]. Wilson et al. haben 2005 gezeigt, dass Makrolid-Antibiotika einen speziesübergreifenden Bindemechanismus an der großen Untereinheit aufweisen [138]. Durch diese Ergebnisse ergibt sich die Frage, ob ein solches Bindemuster auch mit anderen Methoden als einer Kristallstruktur gefunden werden kann.

Als Analyse neben der Kristallstruktur steht in der Molekularbiologie zur Aufklärung von Molekülen in der Bestimmung der Primärsequenz durch Sequenzierung. Mit Hilfe der so ermittelten Sequenzen ist eine Einordnung des untersuchten Moleküls möglich, so können über die Primärsequenz zum Beispiel einzelne Strukturelemente anhand der spezifischen Aminosäurekomposi-

tion in der Primärsequenz identifiziert werden [132, 120]. Anhand solcher Primärsequenzen ist auch eine phylogenetische Zuordnung möglich [49]. Für die Untersuchung der Primärsequenz steht dazu meist die Analyse mit Hilfe von so genannten Sequenzalignments zur Verfügung. Hier werden die Primärsequenzen auf Ähnlichkeiten mit schon genauer analysierten anderen Sequenzen hin untersucht.
Diese Alignments können dann mit Hilfe der MI (Kapitel 2.2.2) bioinformatisch untersucht werden. Dabei werden mit der MI eventuell langreichweitige Korrelationen in Symbolsequenzen untersucht und so Koevolution quantifiziert [91, 53]. Dabei wird in einem Sequenzalignment der Informationsgehalt einer Position über eine andere gemessen und so eine korrelierte Änderung (z.B. Mutation) der einen Spalte in Abhängigkeit der anderen Spalte beschrieben. Die MI hat dabei die Eigenschaft, dass sie 0 ist, wenn keine Änderung stattfindet oder die Änderungen zufällig geschehen. Für die Aufklärung von Koevolution innerhalb des Ribosoms wurden Alignments der ribosomalen Moleküle berechnet und mit Hilfe der MI analysiert. Bei der MI-Berechnung der Alignments treten Sequenzabschnitte oder auch einzelne Residuen auf, die nicht an die anderen Sequenzen aligniert werden können. Diese Bereiche werden vom Alignment-Algorithmus mit Lücken, so genannten Gaps, gefüllt. Biologisch handelt es sich dabei um Sequenzabschnitte, die innerhalb eines Organismus zum Beispiel im Laufe der Evolution in die genomischen Sequenz eingefügt (inseriert) oder aus der genomischen Sequenz entfernt (deletiert) wurden. Die Existenz von Deletion und Insertion erschwert die Interpretation von Gaps innerhalb der MI-Berechnung jedoch, da nicht unterschieden werden kann, durch welches der beiden Phänomene (Insertion oder Deletion) das Gap entstanden ist. Dieser interpretatorischen Schwierigkeit wird in den Berechnungen der MI bisher damit begegnet, dass Alignment-Spalten mit einem Gap-Gehalt von mehr als 20 % in der Berechnung oder der anschließenden Diskussion nicht berücksichtigt werden [22, 91, 53]. In dieser Arbeit wurden hingegen vier unterschiedliche Berechnungsverfahren für die informationstheoretisch korrekte Modellierung der Gaps untersucht.

3.2 Methoden

3.2.1 Alternative Berechnung der *Mutual Information*

Hier werden Verfahren gezeigt, um das Gap-Symbol aus den MI-Berechnungen zu eliminieren. Als Berechnungsgrundlage dient jeweils eine der beiden vorgestellten Gleichungen (2 oder 3). Um der Schwierigkeit zu begegnen das Gap-Symbol zu interpretieren, wurde der Algorithmus zur Berechnung der MI angepasst. Auf Berechnung der MI durch den unmodifizierten Algorithmus wird im Folgenden mit der Bezeichnung ORMI (ORiginal-MI) hingewiesen.

3.2.1.1 SUMI

In dieser neuen Berechnung wird die MI aus der Entropie wie in Gleichung 2 berechnet. Lediglich die Ermittlung der Häufigkeiten $p(x)$, $p(y)$, $p(x,y)$ wird hier modifiziert: es wird zunächst für das aktuell betrachtete Spaltenpaar (X,Y) die maximale Untermenge ausgewählt, die keine Gaps enthält. Diese Berechnung wird im Folgenden als SUbset-MI (SUMI) bezeichnet. Die Wahl der maximalen Untermenge geschieht durch Verwerfen von Paaren, in denen mindestens ein Gap vorhanden ist. Entsteht dadurch eine leere Menge, oder ein Subset, das weniger als

drei Paare enthält, so wird die MI zwischen den beiden Spalten X und Y auf Null gesetzt. Das Schema in Abbildung 3 veranschaulicht die Bildung der Untermenge an zwei Spalten eines Beispielalignments.

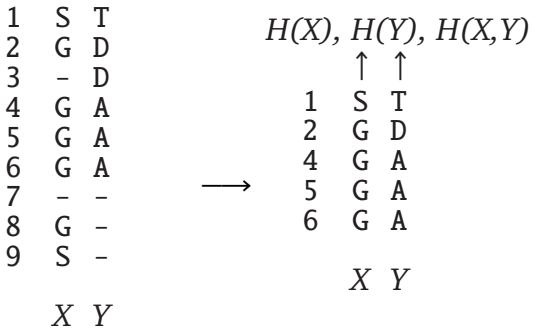

Abbildung 3: Die Bildung der Untermenge vor der Berechnung der SUMI der Alignmentspalten X und Y. Links sind die Spalten X und Y vor der Bildung der Untermenge gezeigt und rechts danach.

Durch dieses Verfahren werden in Spalten mit hohem Gap-Anteil die Gaps nicht wie in der ORMI, (Kapitel 2.2.2) naiv als 21stes Symbol behandelt, sondern mit dem größtmöglichen kontinuierlichen Informationsgehalt gefüllt, der zur Verfügung steht.

3.2.1.2 DEMI

Dieses Verfahren verwendet die Berechnung über Entropien (Gleichung 2) und versucht die Aussagekraft für Spalten mit Gaps so zu maximieren, dass für die Bestimmung der Einzel-Entropien $H(X)$ und $H(Y)$ die Frequenzen aller Realisierungen eingehen, die kein Gap sind. Sie wird als Delta-Entropie-MI (DEMI) bezeichnet, da die Frequenzen mit unterschiedlichen Untermengen des zur Verfügung stehenden Alignments bestimmt werden. Das Schema in Abbildung 4 veranschaulicht das Verfahren der DEMI an dem Beispielalignment.

3.2.1.3 ESMI

Das zuletzt vorgestellte Verfahren verwendet zur Berechnung der MI die Gleichung 3 und wird im Folgenden EnhancedSubset-MI (ESMI) genannt. Dabei werden die relativen Häufigkeiten $p(x)$ und $p(y)$ so gebildet, dass zwar nur die Frequenzen der im Alphabet enthaltenen Symbole (ohne Gap-Symbol) bestimmt werden, für die Gesamtzahl der Sequenzen die Gap-Symbole jedoch mitgezählt werden:

$$p(x) = \frac{\text{Anzahl des Symbols}}{\text{Anzahl aller Symbole mit Gap}}$$

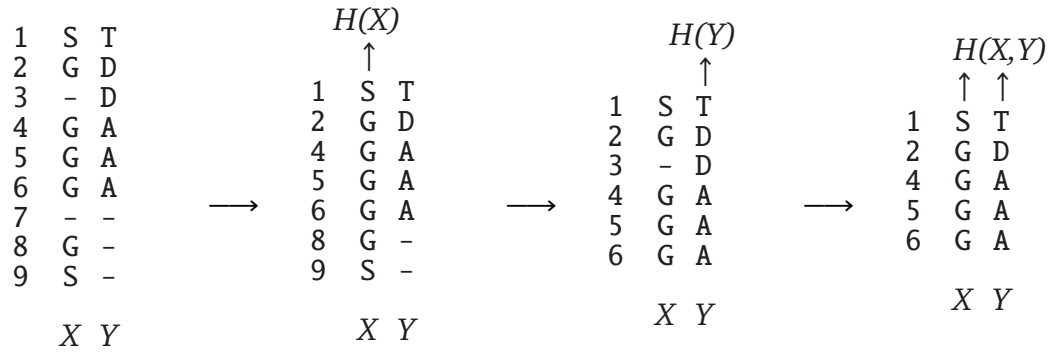

Abbildung 4: Die Bildung der Untermengen vor der Berechnung der DEMI der Alignmentspalten X und Y. Von links nach rechts sind die Untermengen des verwendeten Alignments für die Berechnung von $H(X)$, $H(Y)$ und von $H(X,Y)$ gezeigt.

Bei der ESMI werden für die Frequenzbestimmung also alle beobachteten Symbole verwendet, dabei wird aber berücksichtigt, dass nicht alle möglichen Positionen des Alignments besetzt sind. Das Schema in Abbildung 5 veranschaulicht diese Methode.

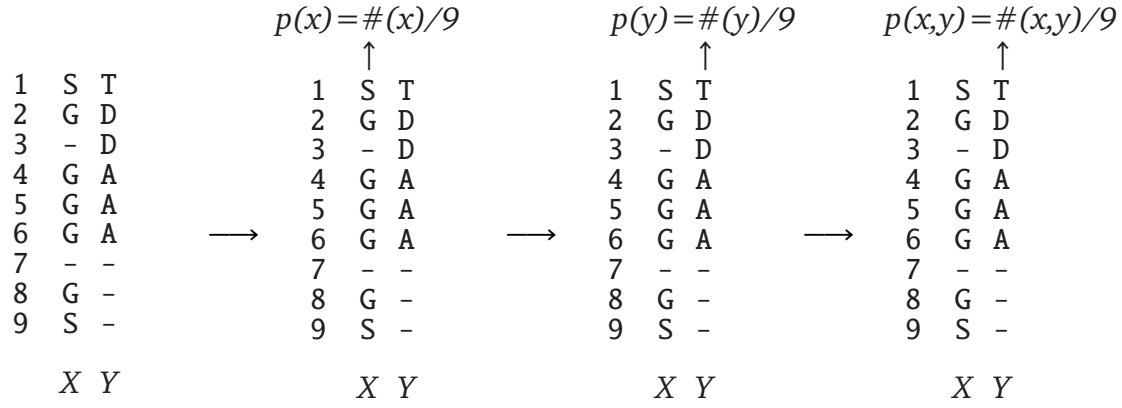

Abbildung 5: Die Berechnung der ESMI für ein Spaltenpaar X und Y eines Beispiel-Alignments. Von links nach rechts sind die Berechnungen der Frequenzen aus den Spalten gezeigt (# = Anzahl). Die Frequenzen für $p(x)$, $p(y)$ und $p(x,y)$ werden jeweils für alle Standardaminosäuren, ohne das Gap-Symbol berechnet.

3.2.2 Normierungen der MI

Da die MI ein statistisches Maß darstellt, das abhängig von den zugrunde liegenden Alignments ist, wurden verschiedene Normierungsverfahren entwickelt, die Hintergrundsignale wie phylogenetische Effekte oder statistisches Rauschen aus den Berechnungen filtern sollen. Phylogenetische Effekte entstehen zum Beispiel durch die Überrepräsentation eines häufig sequenzierten gegenüber eines kaum sequenzierten Organismus. Dadurch kann der untersuchte Datensatz redundante Sequenzen enthalten, wodurch die MI hauptsächlich Signale des überrepräsentierten

Organismus zeigt. Statistisches Rauschen kann durch den so genannten *finite-size* Effekt verursacht werden, der bei einem zu kleinen Datensatz auftritt. Durch die geringe Stichprobengröße kann dann ein zufälliges Signal nicht von einem signifikanten unterschieden werden. Vier dieser Normierungsverfahren werden hier im Speziellen vorgestellt.

3.2.2.1 *Average Product Correction* (APC)

Das von Dunn et al. [40] vorgestellte Verfahren soll den *finite-size* Effekte unterdrücken. Mathematisch wird diese Normierung wie folgt berechnet:

$$APC_{ij} = \mathrm{MI}_{ij} - \frac{\widetilde{\mathrm{MI}}_i \cdot \widetilde{\mathrm{MI}}_j}{\widetilde{\mathrm{MI}}} \tag{7}$$

Dabei sind $\widetilde{\mathrm{MI}}_i$ und $\widetilde{\mathrm{MI}}_j$ jeweils die mittlere MI der MI-Matrixspalten i und j ohne die Diagonalelemente MI_{ii} und MI_{jj}; $\widetilde{\mathrm{MI}}$ stellt die mittlere MI der gesamten MI-Matrix dar.

3.2.2.2 *Row Column Weighting* (RCW)

Die Normierung mit Hilfe der RCW, vorgestellt von Gouveia-Oliveira et al. [56], stellt die Kompensation phylogenetischer Redundanz der Datensätze in den Vordergrund. Die Normierung wird wie folgt berechnet:

$$RCW_{ij} = \frac{\mathrm{MI}_{ij}}{\widetilde{\mathrm{MI}_{ij}}} \quad \text{, mit } \widetilde{\mathrm{MI}_{ij}} = \frac{\sum_{l \neq i} \mathrm{MI}_{il} + \sum_{i \neq k} \mathrm{MI}_{kj} - 2\mathrm{MI}_{ij}}{2(n-1)}, \tag{8}$$

wobei n die Anzahl der Sequenzen im Alignment kennzeichnet.

3.2.2.3 Normierung über die gemeinsame Entropie

Ein weiteres Verfahren, um das statistische Rauschen zu unterdrücken, wurde von Gloor et al. [53] vorgestellt. Dabei wird der beobachtete MI-Wert auf die gemeinsame Entropie der betrachteten Alignmentspalten normiert. Mathematisch kann man diese Normierung wie folgt ausdrücken:

$$\widehat{\mathrm{MI}}_{ij} = \frac{\mathrm{MI}_{ij}}{H_{ij}} \tag{9}$$

3.2.2.4 Normierung über die Spaltenentropie

Um den Einfluss der Spaltenentropie zu eliminieren, verwenden wir folgende weitere Normierung:

$$\widehat{\mathrm{MI}}_{ij} = 1 - \frac{H_{ij}}{H_i + H_j} \tag{10}$$

3.2.2.5 Z-Scores

Um die Signifikanz von einzelnen Werten innerhalb einer Verteilung zu ermitteln wurden die von Gloor et al. [53] vorgeschlagenen Z-Scores auch für jede der eingeführten MI-Berechnungen durchgeführt. Dabei dient das in dieser Arbeit vorgestellt Nullmodell (siehe Kapitel 2) als zugrunde liegende Verteilung. Die Z-Scores sind definiert als Anzahl an Standardabweichungen, die der Wert vom Mittel der zugrunde liegenden Verteilung abweicht. Für die berechneten MI-Werte werden die Z-Scores wie folgt berechnet:

$$Z_{ij} = \frac{\mathrm{MI}_{ij} - \widetilde{\mathrm{MI}}_{ij}}{\sqrt{var(\widetilde{\mathrm{MI}}_{ij})}} \tag{11}$$

Dabei stellt $\widetilde{\mathrm{MI}}_{ij}$ den mittleren Wert der MI aus den Berechnungen des Nullmodells für das aktuelle Spaltenpaar dar und $var(\widetilde{\mathrm{MI}}_{ij})$ die Varianz.

3.2.3 Hidden-Markov-Modelle (HMM)

Hidden-Markov-Modelle (HMM) sind statistische Modelle zur Untersuchung von linearen Sequenzen. Die Anwendung von HMM in der Biologie geht auf Churchill [29] zurück. Die grundlegende Theorie ist die Interpretation einer biologischen Sequenz als Wahrscheinlichkeit einer bestimmten Abfolge von Symbolen. Dabei handelt es sich im Speziellen um entweder das Basen-Alphabet oder das Standardaminosäure-Alphabet. Das folgende Schaubild zeigt die Struktur eines HMM.

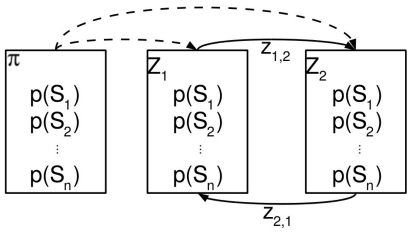

Abbildung 6: Struktureller Aufbau eines Hidden-Markov-Modells. Im Schaubild sind der Anfangszustand π und die Zustände Z_1 und Z_2 des HMMs durch Kästen veranschaulicht. In den Kästen sind die Wahrscheinlichkeiten der jeweiligen Symbole $p(S_1)$ bis $p(S_n)$ gezeigt. Diese Emissionswahrscheinlichkeiten können in den unterschiedlichen Zuständen verschieden sein. Die Pfeile zeigen die Übergangswahrscheinlichkeiten $z_{1,2}$ und $z_{2,1}$ zwischen Z_1 und Z_2.

Wie Abbildung 6 verdeutlicht, wird aus einem Anfangszustand π mit einer Anfangsverteilung das erste Symbol erzeugt und die Sequenzgenerierung gestartet. Die Anfangsverteilung im Zustand π, genau so wie die Übergangswahrscheinlichkeit von π nach Z_1 oder von π nach Z_2, werden meist gleichverteilt. Diese Verteilungen werden über das Training des Modells an einem realen Trainingsdatensatz variiert, sodass sie das reale Beispiel abbilden. Die Symbole werden dann nach den jeweiligen Verteilungen in Z_1 oder Z_2 emittiert. Nach jedem Symbol wird mit den über die Übergangswahrscheinlichkeiten $z_{1,2}$ und $z_{2,1}$ bestimmt, wie wahrscheinlich das nächste Symbol aus der Verteilung des jeweils anderen Zustands gewählt wird, dabei ist die Verweilwahrscheinlichkeit genau $1 - z_{1,2}$ oder $1 - z_{2,1}$. Mit solchen HMMs lassen sich zum Beispiel CpG-Inseln identifizieren oder synthetische Sequenzen zu CpG-Inseln generieren [30].

Aus empirischen Sequenzdaten lassen sich durch Beobachtung alle Wahrscheinlichkeiten eines solchen HMM bestimmen. Diese HMM mit den beobachteten Parametern sind für verschiedene einzelne Proteine oder ganze Proteinfamilien in der Pfam Datenbank (Pfam-DB) [47, 125] enthalten. Aus diesen parametrischen Beschreibungen von Proteinfamilien kann algorithmisch die Wahrscheinlichkeit bestimmt werden, dass eine vorliegende Sequenz zu dieser Proteinfamilie gehört. Mit dem `HMMER`-Algorithmus [46] kann die Sequenz eines Proteins in einer AS-Sequenz identifiziert werden. Dabei wird diese gleichzeitig mit einem *E*-Value bewertet, der beschreibt wie hoch die Wahrscheinlichkeit ist, dass das identifizierte Teilstück zufällig in der AS-Sequenz enthalten ist. In dieser Arbeit wurde `HMMER2.0` in der Version 2.3.2 verwendet.

3.2.4 Sequenzidentifikation

Da es für viele der ribosomalen Proteine nicht genügend experimentelles Datenmaterial gibt, wurden die Sequenzen aus Genomdaten mit Hilfe von *hidden markov* Modellen (HMM) extrahiert (Kapitel 3.2.3), sodass ein ausreichend großer Datensatz zur Verfügung steht. Die mit den HMM identifizierten Sequenzen dienen als Grundlage für die MI-Untersuchungen (Kapitel 2.2.2). Die Datengrundlage für den `HMMER2.0`-Algorithmus wurde aus der GenBank Datenbank (GB-DB) des *National Center for Biotechnology Information* (NCBI) [13] extrahiert. Dazu wurden Genomsequenzen von Bakterien aus der GB-DB automatisiert heruntergeladen und lokal weiter prozessiert.

Alle gefundenen Genomsequenzen wurden anschließend mit Hilfe des `Biophython`-Moduls [32] in alle möglichen Proteinsequenzen translatiert. In den so erhaltenen genomischen Proteinsequenzen wurden die Sequenzen der ribosomalen Proteinen mit dem `HMMER2.0`-Algorithmus [46, 125] identifiziert. Für die Identifikation der Proteinsequenzen wurde die Qualität der gefundenen Teilsequenz über den *E*-Value des `HMMER2.0`-Algorithmus bestimmt. Dabei geben die *E*-Values Auskunft über die Wahrscheinlichkeit dafür, dass der Algorithmus eine Sequenz identifiziert hat, die zufällig zu dem HMM passt. Da für einige HMM nur ein sehr kleiner Datensatz zur Verfügung stand, wurde ein *E*-Value Schwellenwert von 10^{-5} gewählt. Durch diesen Filter wird die Identifikation falsch positiver Sequenzen verringert. Ein weiteres Prozessieren der Sequenzen war notwendig, da der `HMMER2.0`-Algorithmus keine Stop-Codons erkennt und dadurch auch Proteinsequenzen identifiziert, die innerhalb der Sequenz ein Stop-Codon aufweisen. Aus diesem Grund wurden die Sequenzen verworfen, deren Länge durch ein Stop-Codon um mehr als 20 % verkürzt wurde.

Für die Identifikation der rRNA wurden keine HMMs verwendet, da die Sequenzen entweder als Eintrag in der GenBank oder in den Genomsequenzen annotiert vorlagen. Daher wurden die Sequenzen der 16S rRNA separat für jeden Organismus aus der GenBank Datenbank heruntergeladen und die Sequenzen der 5S und 23S rRNA aus den genomischen Daten extrahiert.

Die Sequenzen, die für das Alignment bereitstanden, stammen aus den genomischen Daten von 357 Bakterien-Spezies. Die genomischen Sequenzen dieser Bakterien wurden aus der GB-DB [14] heruntergeladen, dabei standen für die unterschiedlichen Organismen verschieden viele Datensätze zur Verfügung, sodass hier 778 genomische Sequenzen auf ribosomale Moleküle hin untersucht wurden. Dieser Unterschied in Sequenzen und Organismen ergibt sich aus der unterschiedlichen Relevanz der einzelnen Organismen für die Forschung und die daraus resultierende unterschiedlich häufige Sequenzierung.

3.2.5 Netzwerkanalyse

Um einen komplexen Zusammenhang in einem großen System zu identifizieren werden vielfach Netzwerkanalysen durchgeführt [84, 104, 20]. Aus unseren koevolutionären Signalen der MI ergeben sich solche Netzwerke für die starke Koevolution von AS bzw. Nukleotiden. Generell kann bei diesen Netzwerkanalysen zwischen zwei großen Bereichen unterschieden werden. Zum Einen gibt es die Analyse dynamischer Prozesse innerhalb des zu untersuchenden Netzwerkes. Zum Anderen ist es möglich, verschiedene topologische Kenngrößen eines Netzwerkes zu bestimmen. Da es in der vorliegenden Arbeit um die Analyse von Proteinstrukturen geht, wird die dynamische Netzwerkanalyse nicht verwendet und das Augenmerk auf die Kenngrößen der Topologie gelegt.

Die zu untersuchenden Graphen[4], sind nicht aus den strukturellen Informationen des physikalischen Abstandes der AS innerhalb der Proteine entstanden, sondern aus den berechneten MI-Matrizen generiert. Dazu wurden aus den berechneten MI-Matrizen Distanzmatrizen erstellt, die sich aus folgendem Zusammenhang ergeben:

$$D = max\,(MI) - MI \tag{12}$$

Da sich topologische Unterschiede von ungewichteten Graphen nur bei nicht vollständiger Vernetzung zeigen, stellt sich als nächstes die Frage, welche Distanzen als Kontakt gewertet werden und welche nicht. Dazu wurden alle MI-Matrizen in Graphen umgewandelt, deren Vernetzung hier von 5 % bis 100 % der maximalen, d.h. kompletten Vernetzung, variiert. Dieser Vernetzungsgrad beschreibt die Dichte des Netzwerks. So wurden aus jeder MI-Matrix 20 Netzwerke mit Dichten zwischen 5 % und 100 % in jeweils 5 % Schritten erstellt. Bei den so entstandenen Graphen handelt es sich um sogenannte ungerichtete Graphen, das heißt die Verbindung zweier Knoten[5] durch eine Kante hat keine Richtung, unterscheidet also nicht, ob die Kante von Knoten v_i zu Knoten v_j zeigt oder umgekehrt. Von diesen Netzwerken wurden anschließend die folgenden Kenngrößen bestimmt:

[4] Das Netzwerk wird hier durch einen Graphen repräsentiert.
[5] Knoten sind hier die Repräsentation der Aminosäuren oder Nukleotide.

Die mittlere Pfadlänge (l) ist die mittlere Verbindungslänge zwischen zwei Knoten v_i und v_j des Graphen, also die durchschnittliche Anzahl der Kanten, die auf dem Weg von einem Knoten v_i zu einem beliebigen anderen Knoten v_j passiert werden. Für die mittlere Pfadlänge werden alle vorhandenen Pfadlängen bestimmt und über die Gesamtzahl der Pfade normiert.

Der Grad (k) beschreibt die Anzahl an Kanten e eines Knotens v_i.

Die *Closeness* (c) bestimmt die Zentralität eines Knotens. Die *Closeness* wird für jeden Knoten v_k eines Graphen mit N Knoten ermittelt als den reziproken Wert der mittleren kürzesten Pfadlänge g zu allen anderen Knoten v_i des Netzwerkes. Handelt es sich dabei um einen ungerichteten Graphen wird die Anzahl der Knoten gezählt, die auf dem Weg von v_i zu v_k passiert werden [50]. Damit ergibt sich für die Closeness folgender mathematischer Ausdruck:

$$c_{v_k} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N} g(v_i, v_k)}, \text{ mit } v_i \neq v_k \tag{13}$$

Die *Betweenness* (b) ist ein anderes Maß für die Zentralität eines Knotens v_k. Sie wird bestimmt durch die Anzahl der kürzesten Pfade g, die von jedem Knoten v_i zu einem anderen Knoten v_j über den Knoten v_k führen. Diese Summe wird auf die Anzahl aller Pfade zwischen v_i und v_j normiert, sodass sich folgender Ausdruck ergibt:

$$b_{v_k} = \sum_{i,j} \frac{g_{v_i v_k v_j}}{g_{v_i v_j}}, \text{ mit } v_i \neq v_j,\ v_i \neq v_k,\ v_j \neq v_k \tag{14}$$

Die Clusterzugehörigkeit (cm) wird als Nummer des Clusters angegeben, zu dem der beobachtete Knote v_i gehört. Ein Cluster ist dabei definiert als zusammenhängendes Netzwerk. Existieren nur Knoten und keine Kanten in einem Netzwerk so bildet jeder Knoten für sich einen eigenen Cluster.

Die Dichte (d) beschreibt die Anzahl der vorhandenen Kanten e im Verhältnis zu allen möglichen Kanten E. Dieses Netzwerkmaß repräsentiert also den Verknüpfungsgrad des Graphen und wurde als Kontrollgröße mitberechnet. Die Dichte des Netzwerkes ist somit definiert als:

$$d = \frac{e}{E} \tag{15}$$

Die Gradverteilung (dk) ist die Verteilung der im Graphen beobachteten Grade k aller Knoten v.

Grad-Grad-Korrelation Dieses Maß berechnet über alle M Kanten eines Netzwerkes die Korrelation der Grade k. Dabei werden die an den Knoten v_j und v_l, verbunden durch die Kante e_i, gefundenen Grade korreliert. Die Grad-Grad-Korrelation stellt also so etwas wie die Pearson-

Korrelation der Grade k an den verbundenen Knoten j und l dar und wird nach folgendem Zusammenhang berechnet [101]:

$$r = \frac{M^{-1}\sum_i j_{e_i} l_{e_i} - \left[M^{-1}\sum_i \frac{1}{2}(j_{e_i} + l_{e_i})\right]^2}{M^{-1}\sum_i \frac{1}{2}(j_{e_i}^2 + l_{e_i}^2) - \left[M^{-1}\sum_i \frac{1}{2}(j_{e_i} + l_{e_i})\right]^2} \tag{16}$$

Zufällige Netzwerke wurden generiert, um die Netzwerkmaße der realen Netzwerke einzuschätzen. Dazu wurden die Netzwerkmaße der realen Netzwerke mit denen zufälliger Netzwerke gleicher Dichte verglichen. Die zufälligen Netzwerke wurden durch das Einfügen von Kanten generiert, wobei die Wahrscheinlichkeit, dass zwischen zwei beliebigen Knoten eine Kante existiert, der Dichte entspricht. Für den Vergleich der Netzwerkmaße wurden zufällige Netzwerke unterschiedlicher Größe (10, 50, 100, 150 und 200 Knoten) analysiert.

Die Netzwerkmaße wurden mit Hilfe des `igraph`-Paketes [33] und der Statistiksoftware `R` [113] berechnet.

3.2.6 Evolutionsmatrizen

Um einen Überblick über die koevolutionären Muster innerhalb des Ribosoms zu erhalten wurden das maximale und das mittlere Signal der Z-Scores für jede Proteinpaarung bestimmt und in eine Z-Matrix geschrieben. Um eine Normierung auf die Effekte aus Kapitel 3.2.2 einzubeziehen, arbeiten wir hier mit diesen Z-Matrizen statt der originalen MI-Matrix. Diese Z-Matrizen wurden anschließend in Distanzmatrizen umgewandelt (Gleichung 12) und mit Hilfe von `hclust` [99] nach Ähnlichkeit geordnet. Diese Matrizen geben damit also einen Überblick über die Ähnlichkeit des Moleküls bezüglich des maximalen oder mittleren koevolutionären Signals zu allen anderen Residuen.

3.2.7 Bestimmung der Antibiotika-Bindestellen

Angeregt durch die Publikation von Wilson et al. [138], die einen speziesübergreifenden Bindemechanismus vom Makrolid-Antibiotika beschreibt, wurden die Z-Score-Matrizen für alle inter- und intramolekularen MI-Berechnungen bestimmt. Aus den Kristallstrukturen zweier Makrolid-Antibiotika Telithromycin (PDB-Code 3OI3) und Azithromycin (PDB-Code 3HOZ) wurden die Bindepositionen der Antibiotika bestimmt. Dafür wurde ein Interaktionsradius von 13 Å um das Antibiotikum in den PDB-Strukturen angenommen und alle Residuen innerhalb dieses Radius wurden als potentialle Interaktionspartner definiert. Beide ribosomalen Strukturen sind aus *T. thermophilus*. Da sich in der HIV-1 Protease Resistenzmuster mit Hilfe der MI finden ließen [62, 81], vermuten wir hier, dass sich eventuell auch die Positionen der Antibiotika wiederfinden könnten.

3.3 Resultate

3.3.1 Ribosomale Mutual Information

Tabelle 1: Absolute Häufigkeiten von Organismen die für die jeweilige Anzahl an Genomsequenzen gefunden wurde.

Anzahl Genomsequenzen	1	2	3	4	5	6	7	8	9	10	11	12
Anzahl Organismen	131	177	16	12	2	6	3	2	1	1	1	1
Anzahl Genomsequenzen	15	16	19	22								
Anzahl Organismen	1	1	1	1								

Die Tabelle 1 zeigt, dass sich für den Hauptteil der Organismen bis zu drei oder vier Genomsequenzen finden lassen. Es sind aber auch Organismen zu finden, von denen weit mehr Genomsequenzen in der GenBank-DB zu finden sind. Die vier Bakterienstämme, von denen am meisten Genomsequenzen gefunden wurden sind *Streptococcus pyogenes* mit 15 Genomsequenzen, *E. coli* mit 16, *Prochlorococcus marinus* mit 19 und *Staphylococcus aureus* mit 22 Sequenzen. Die ungleichmäßige Verteilung der Genomsequenzen führt zu einer unterschiedlich starken Repräsentation der Moleküle innerhalb der Alignments. Um diesen Einfluss abschätzen zu können, wurde aus dem Alignment der 16 S rRNA ein phylogenetischer Baum berechnet und überprüft.

3.3.1.1 Einfluss des Alignments

Die Berechnung der MI hängt stark von dem zugrunde liegenden Alignment ab, daher wurde zur Kontrolle des Alignment-Algorithmus `clustalw` [130, 28, 78] das Alignment der 16 S rRNA in einen phylogenetischen Baum umgerechnet (Abbildung 7) und mit dem phylogenetischen Baum von Wu et al. [143] verglichen. Für die Berechnung des Alignments und des phylogenetischen Baumes wurden die Algorithmen `clustalw` und `Phylip` [45] mit den Standardparametern verwendet. Der Vergleich der Bäume zeigte eine gute Übereinstimmung der Speziescluster, sodass die Standardparameter von `clustalw` bei allen Alignments in dieser Arbeit verwendet wurden.

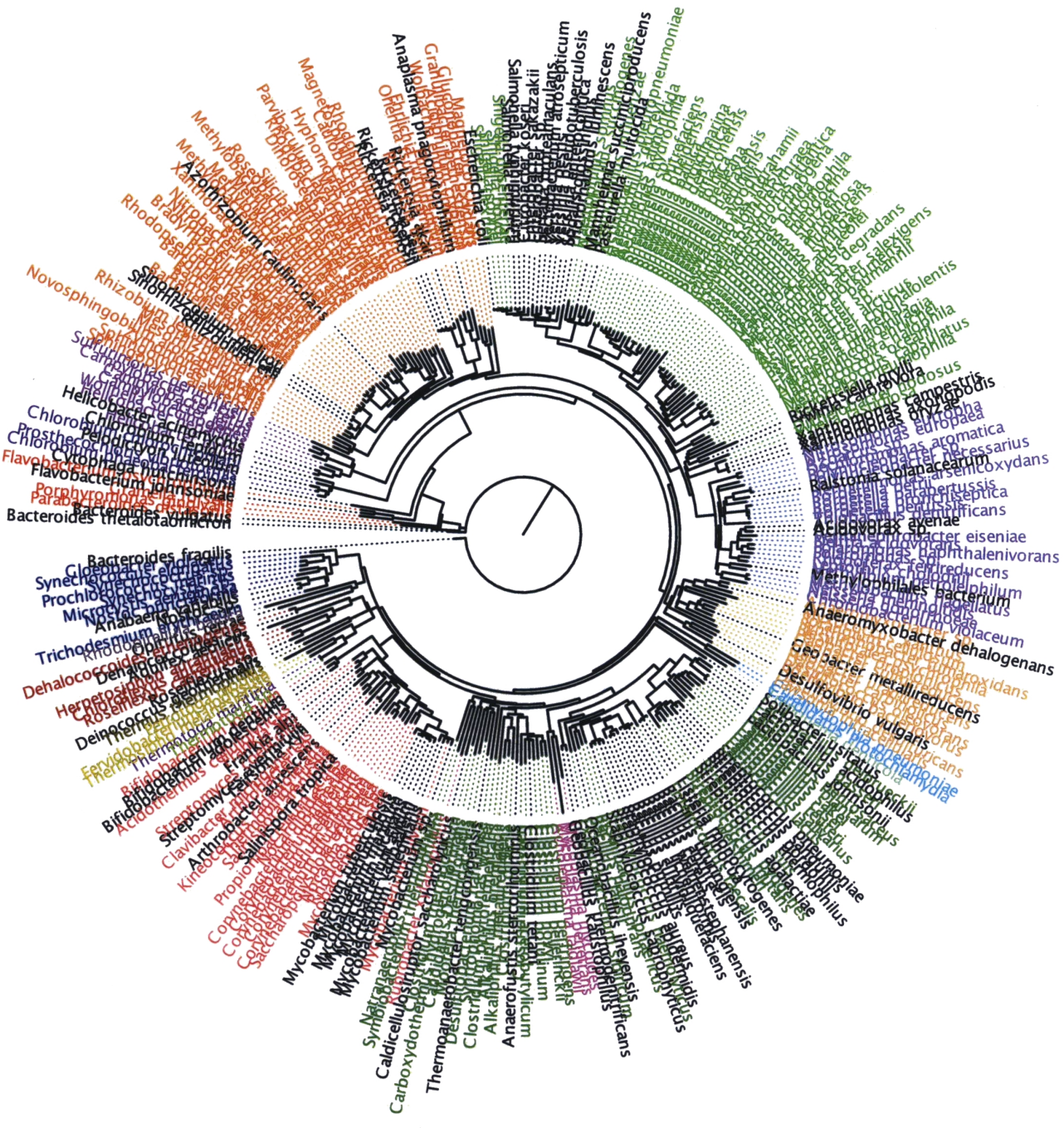

Abbildung 7: Phylogenetischer Baum der 16S rRNA der in dieser Arbeit verwendeter bakteriellen Spezies. Die Taxa sind nach bakteriellen Klassen koloriert. Die Kolorierung orientiert sich dabei an der von Wu et al.. Nicht kolorierte Taxa (schwarz) sind nicht im phylogenetischen Baum von Wu et al. enthalten.

3.3.1.2 Einfluss der Stichprobengröße

Eine weitere Einflussquelle bei der MI-Berechnung ist die Stichprobengröße. Durch die unterschiedliche großen Trainingsdatensätze wurden in den Genomsequenzen der Bakterien für die unterschiedlichen ribosomalen Moleküle verschieden viele Sequenzen durch den `HMMER`-Algorithmus identifiziert. Dadurch stehen für die Berechnung der Alignments in den unterschiedlichen Molekülen zum Teil unterschiedlich viele Sequenzen für die Berechnung der MI zur Verfügung. Der Einfluss der Sequenzanzahl auf zufällig erzeugte MI soll mit Hilfe von Abbildung 8 veranschaulicht werden. Dabei wurde die Anzahl der zufällig erzeugten AS-Paare variiert und so die verschiedenen Stichprobengrößen simuliert. Die MI wurde für jede Stichprobengröße aus 10 000 Berechnungen gemittelt. Aus Abbildung 8 wird deutlich, dass die MI

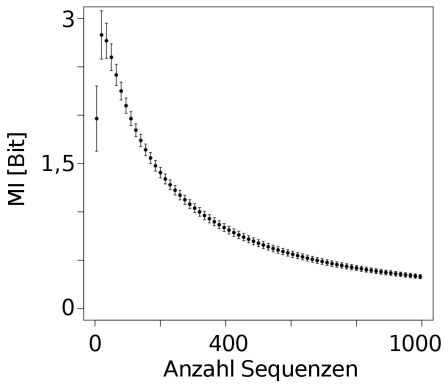

Abbildung 8: MI für zufällige AS-Paare in Abhängigkeit der Anzahl. Aufgetragen ist die berechnete MI gegen die Anzahl der verwendeten Sequenzen. Die Fehlerbalken zeigen die Standardabweichung bei 10 000 Berechnungen mit jeweils der gleichen Anzahl von AS-Paare.

selbst bei zufällig erzeugten AS-Paare sehr hoch ist, da bei kleiner Stichprobengröße die MI für zufällig erzeugte AS-Paare sehr nahe dem theoretischen Maximum der MI kommt, das für das Standard-AS-Alphabet bei $\log_2(20) \approx 4{,}321$ liegt. Außerdem ist zu erkennen, dass die statistisch unabhängig erzeugte MI für ein Protein-Alignment erst bei ca. 725 Sequenzen unter 0,5 sinkt. Da in der Biologie häufig Alignments mit einer deutlich geringeren Anzahl an Sequenzen verwendet werden, ist die Verwendung eines Nullmodells um so wichtiger. Ein eigenes Nullmodell ist in Kapitel 2 beschrieben. Dieses Nullmodell wird in allen folgenden Berechnungen angewendet, sodass die hier präsentierten Daten auf der Analyse der Z-Scores der MI-Analyse beruhen.

3.3.1.3 Einfluss der Phylogenie

Ein weiterer Faktor, der die Berechnung der MI beeinflusst, ist die phylogenetische Zusammensetzung des zugrunde liegenden Alignments. Dabei kann es durch die Überrepräsentation einer Spezies oder einer Familie innerhalb des untersuchten Alignments zu einer erhöhten MI kommen, die jedoch nur innerhalb dieser überrepräsentierten Spezies existiert und keine allgemeine Eigenschaft des analysierten Alignments ist. Dieser phylogenetische Effekt wurde auch von Dunn et al. [40] beschrieben. Um diesen Effekt innerhalb des Alignments auf die Berechnung der MI zu beachten wurden mehrere Verfahren entwickelt, die mit Hilfe von Korrekturter-

men [56, 91] oder Normierungsverfahren (vergl. Kapitel 3.2.2), diesem Effekt entgegenwirken sollen. Das in dieser Arbeit vorgestellte Nullmodell (Kapitel 2) berücksichtigt den phylogenetischen Effekt implizit, da eine mittlere Obergrenze für das vorliegende Alignment berechnet wird.

3.3.1.4 Einfluss von Gaps

In Abbildung 9 wurde die MI des Alignment des ribosomalen Proteins L3 mit allen vier Methoden (vergleiche Kapitel 2.2.2) berechnet und verglichen.

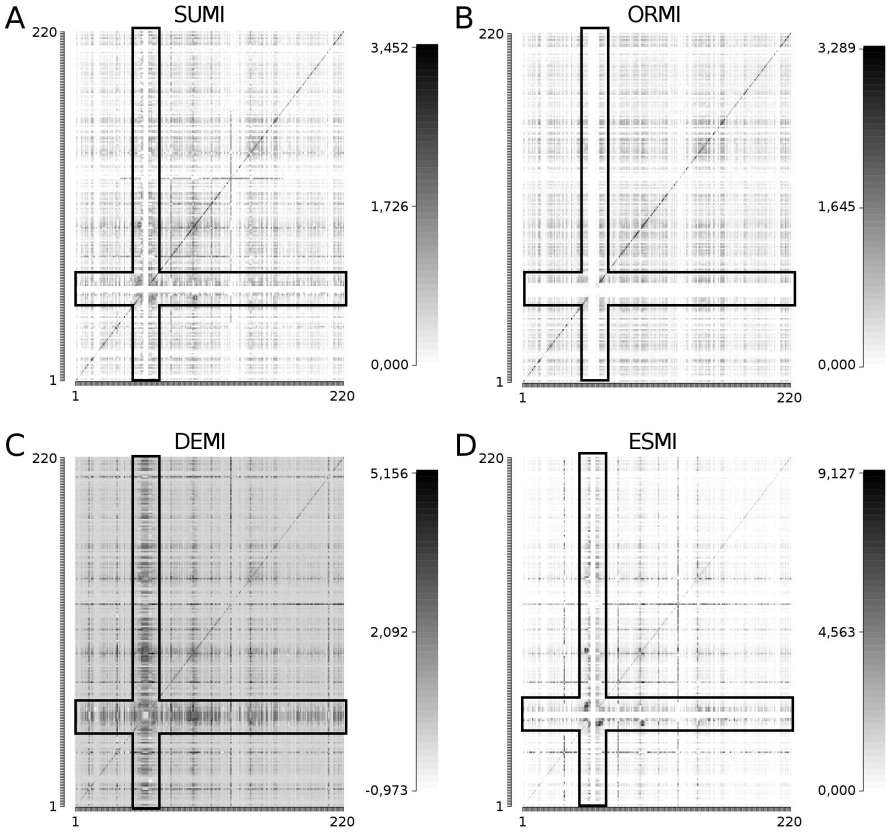

Abbildung 9: Berechnung der MI mit Hilfe der SUMI (A), der ORMI (B), der DEMI (C) und der ESMI (D). Der hervorgehobene Bereich beinhaltet viele Gaps innerhalb des Alignments. Die Farbkodierung der Matrizen zeigt den Informationsgehalt in Bit.

Zu erkennen sind die unterschiedlichen Ergebnisse in den Spaltenpaarungen, die eine hohe Anzahl an Gaps aufweisen (schwarzer Rahmen in Abbildung 9), dabei wird die jeweils höheren MI der modifizierten Methoden in diesem Bereich deutlich. Diese Beobachtung lässt darauf schießen, dass die Modifizierung der ORMI zu stärkeren Signalen in Bereichen mit hohem Gap-Anteil kommt. Des weiteren kann eine unterschiedliche Skala der Werte für vier unterschiedlichen MI-Berechnungen festgestellt werden. Hier ist besonders auffällig, dass die Werte der ESMI (Kapitel 3.2.1.3) ein höheres Maximum erreichen, als durch die theoretische Abschätzung durch Gleichung 2 möglich wäre (vergleiche Abbildung 9 (D)). Dies lässt sich durch die unterschiedlichen Frequenzen der Einzelhäufigkeiten $p(x)$ und $p(y)$ erklären, die in Gleichung 3

durch einen sehr hohen Gap-Gehalt zu sehr kleinen Werten, in Summe auch kleiner 1, führen können. Auch die Ergebnisse der DEMI (Kapitel 3.2.1.2) sind teilweise größer als das theoretische Maximum der MI (vergleiche Abbildung 9 (C)). Bei der DEMI kommt hinzu, dass durch die unterschiedliche Bestimmung der Einzelfrequenzen für die Entropie sogar negative Werte entstehen können. Diese Eigenschaften der ESMI und der DEMI erschweren eine Interpretation der berechneten Werte, daher werden diese beiden Verfahren in dieser Arbeit nicht weiter verwendet. Wie in den Ergebnissen der SUMI und ORMI (vergleiche Abbildung 9 (A) und (B)) deutlich wird, kann die SUMI im Gegensatz zur ORMI in Bereichen, die viele Gaps enthalten, noch interpretierbare Ergebnisse erzielen. Demzufolge werden alle Ergebnisse dieser Arbeit mit Hilfe der SUMI berechnet und deren Signifikanz mit Hilfe des Nullmodells (Kapitel 2) und der Z-Scores (Kapitel 3.2.2.5) bestimmt.

3.3.2 Sensitivitäts-Analyse für die intermolekulare Koevolution

Die Analyse von Sequenz-Alignments mit Hilfe der MI kann nicht nur innerhalb von Molekülen (intramolekular) Aufschluss über koevolutionäre Prozesse geben, sondern auch zwischen zwei Molekülen (intermolekular). Dazu werden die Sequenz-Alignments beider Proteine konkateniert[6]. Bei der Berechnung der MI dieser kombinierten Sequenzen ergibt sich als Ergebnis eine MI-Matrix, die sowohl jeweils die intramolekulare MI von Molekül A und Molekül B als auch die intermolekulare MI zwischen Molekül A und Molekül B enthält. Um dieses Ergebnis zu berechnen, ist es notwendig, dass sich die Sequenzen aus dem gleichen Organismus hintereinander befinden, sich also die Alignment-Reihen aus Sequenzen des gleichen Organismus bilden. Dabei ist zu berücksichtigen, dass in den hier vorliegenden Alignments der ribosomalen Moleküle unterschiedlich viele Sequenzen in Molekül A und in Molekül B für den gleichen Organismus identifiziert wurden. Außerdem kann es vorkommen, dass nicht jedes Molekül für alle Organismen identifiziert wurde. Daher ergibt sich für die Berechnung der intermolekularen MI das Problem, dass in den Sequenz-Alignments der Moleküle nur jeweils die Sequenzen betrachtet werden können, die in der Schnittmenge der Organismen für beide Moleküle vorkommen. Daher wurde für die Berechnung der intermolekularen MI jeweils ein neues „Alignment“ erstellt. Dazu wurde zunächst die Organismen-Schnittmenge bestimmt, für die in beiden Alignments der einzelnen Moleküle Sequenzen vorhanden waren und anschließend wurden, wenn unterschiedlich viele Sequenzen der Moleküle für den gleichen Organismus gefunden wurden, die Sequenzen kombinatorisch aneinander gehängt, sodass jede Sequenz des Moleküls A für Organismus 1 hinter jeder Sequenz des Moleküls B für Organismus 2 steht. Das Alignment wurde nicht neu berechnet, sondern aus dem bestehenden intramolekularen Alignment verwendet.

Diese Methode erlaubt es die intramolekulare MI für jede intermolekulare Paarung mit zu berechnen, sodass eine Einschätzung über die Stärke der Veränderung durch diesen kombinatorischen Ansatz möglich ist. Um den Einfluss dieses Verfahrens zu überprüfen wurde die intramolekulare MI nach jedem Teilschritt (Ursprungs-Alignment (original), Schnittmengenbildung (reduziert), Kombinatorik (vervielfacht) und dem kombinierten Ansatz (kombiniert)) berechnet und anschließend mit der intramolekularen MI des „Originals“ verglichen. Die Korrelationsplots sind in Abbildung 10 am Beispiel der Moleküle L5 und L34 gezeigt, da sich ein tendenziell ähnliches Bild für alle Molekülpaarungen ergeben hat.

[6] Hierbei liegt $p(x)$ in Molekül A und $p(y)$ in Molekül B

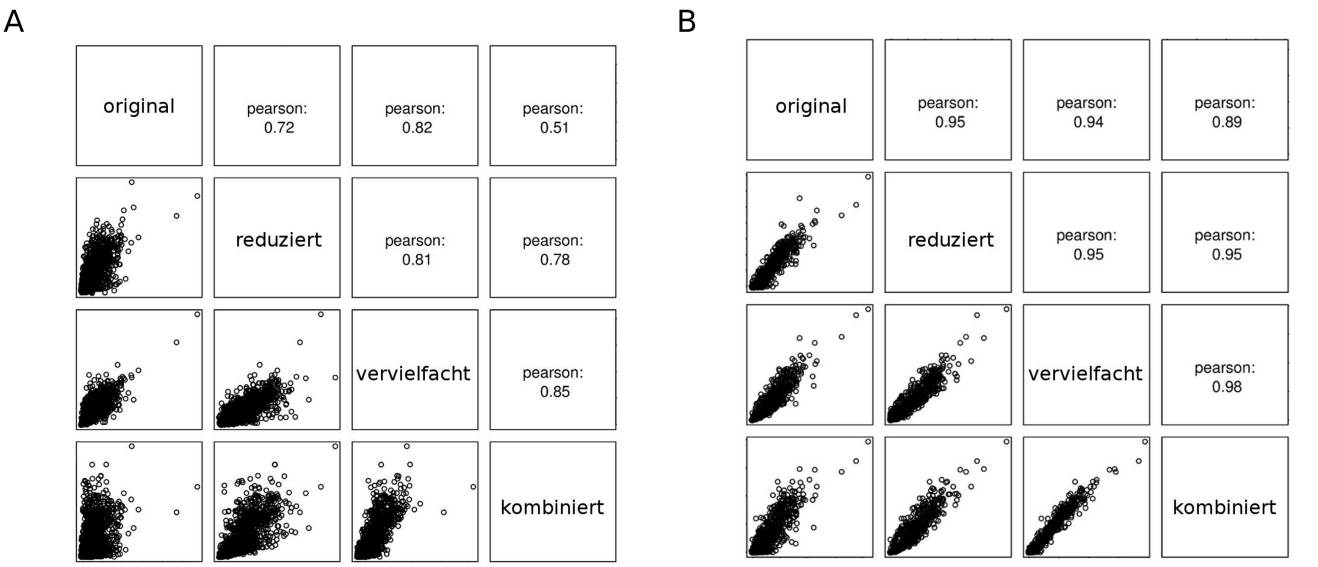

Abbildung 10: Sensitivitäts-Analyse für die intra-MI im intermolekularen Alignment von L5 (A) und L34 (B). Im jeweils unteren Dreieck der Grafenmatrix sind die Scatterplots für den direkten Vergleich des Originals mit allen Zwischenschritten (reduziert und vervielfacht) und dem kombinierten Ansatz (kombiniert). Im oberen Dreieck der Matrix sind jeweils die Pearson-Korrelationskoeffizienten des entsprechenden Vergleichs eingetragen.

Wie in Abbildung 10 zu erkennen ist, werden die Ergebnisse der MI-Berechnung für einzelne Moleküle beeinflusst. Dabei werden die Ergebnisse stärker durch die Kombination (vervielfacht) beeinflusst als durch das Bilden der organismischen Schnittmenge (reduziert). Die Korrelation nach Pearson wurde berechnet, um den Grad der Veränderung vor und nach der oben beschriebenen Methode abzuschätzen. Für jedes Protein entstanden so 50 Vergleiche, nämlich jedes Protein in Kombination mit jedem anderen. Der mittlere Pearson-Korrelationskoeffizient jedes Proteins und die entsprechende Standardabweichung über alle Vergleiche sind in Abbildung 11 zusammengefasst.

Die Zusammenfassung der Sensitivitäts-Analyse in Abbildung 11 zeigt, dass die kombinatorische Methode für die Berechnung des intermolekularen Koevolutionssignals einen Einfluss auf die berechneten Z-Scores hat, der die Korrelation zum „originalen“ Alignment nicht unter 0,5 im Pearson-Koeffizienten sinken lässt. Daher wird der so auftretende Fehler als marginal angenommen und das berechnete intermolekulare Koevolutionssignal nicht mit einem weiteren Korrekturterm verrechnet.

3.3.3 MI und Antibiotika-Bindestellen

Am Beispiel des ribosomalen Proteins L32 und L4 wird die Bindung von Telithromycin detaillierter untersucht (vergl. Abbildung 12), dazu wurden die Matrix-Spalten der Z-Score Matrizen summiert, um abzuleiten, ob diese Residuen intramolekular über besonders viele hohe koevolutionäre Signale verfügen. Des Weiteren wurden die intermolekularen Signale der Z-Score-Matrizen auf besonders hohe Signale hin untersucht.

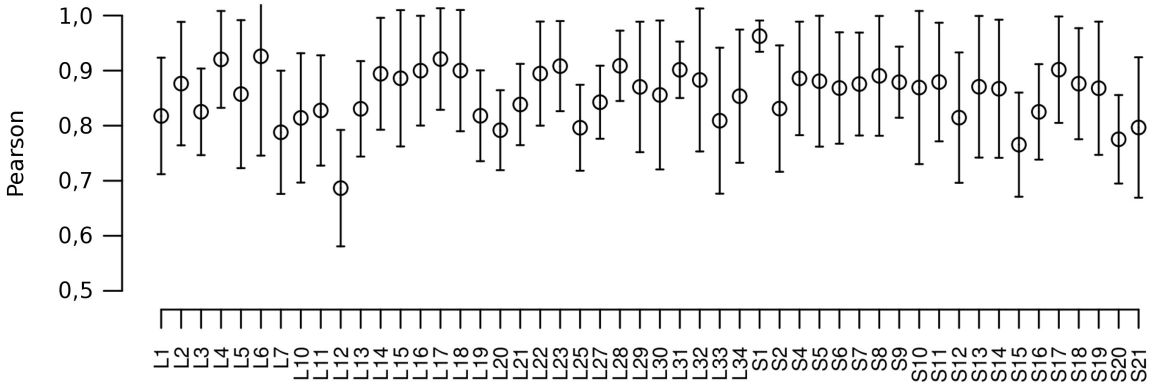

Abbildung 11: Pearson-Korrelation der Z-Scores von intra- und intermolekularen MI-Berechnungen. Dabei sind die mittleren Pearson-Korrelationskoeffiezienten aller Protein-Protein-Vergleiche für jedes Protein aufgetragen. Die Fehlerbalken zeigen die Standardabweichung.

In allen drei Figuren Abbildung 12 (A), (B) und (C) wird deutlich, dass durch einfache Analyse der Z-Score-Matrizen keine Antibiotika-Bindestellen identifizieren lassen, da sich in keiner der Abbildungen die Z-Scores an den Bindestellen hervorheben. Auch die Betrachtung der intermolekularen Z-Scores zeigen keine durch Antibiotika induzierte Koevolution der ribosomalen Proteine.

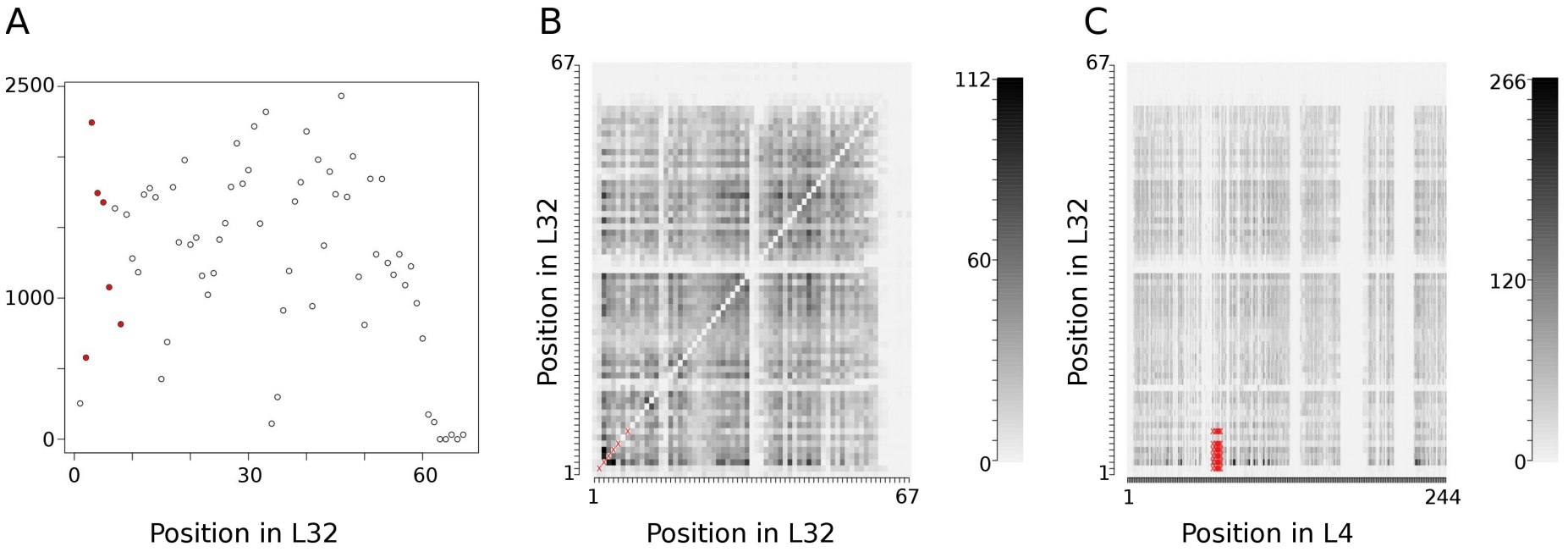

Abbildung 12: Überblick der Z-Score-Matrizen in Verbindung mit den Antibiotika bindenden Residuen. In (A) sind die summierten Z-Score-Matrix Spalten über dem Index des Alignments aufgetragen. (B) zeigt die intramolekulare Z-Score-Matrix des ribosomalen Proteins L32 und (C) die intermolekularen Z-Scores zwischen L32 und L4. Die Antibiotika bindenden Residuen sind jeweils rot markiert.

3.3.4 Ribosomale MI-Netzwerke

Die Analyse globaler Zusammenhänge aus großen Datenmengen und daraus resultierende Ableitung von höher hierarchischen Zusammenhängen wird in der Bioinformatik mit Hilfe von Netzwerkanalysen durchgeführt [23, 84, 11, 20]. Vor Allem die Arbeit von Butte et al. [23] zeigt, dass es möglich ist mit Hilfe von Netzwerken aus MI-Daten biologisch relevante Gencluster zu identifizieren. Ein ähnlicher Ansatz soll hier auf die berechnete Sequenz-MI der ribosomalen Moleküle übertragen werden. Dazu wurden die Z-Score-Matrizen in Distanzmatrizen umgewandelt und so Netzwerke verschiedener Dichten erzeugt, indem ein Z-Score-Schwellenwert festgelegt wurde, bei dem 5 %-100 % der Z-Score-Werte als vernetzt angenommen werden. Die aus diesen Netzwerken errechneten Netzwerkmaße (vergleiche Kapitel 3.2.5) wurden mit denen aus zufällig erzeugten Netzwerken der gleichen Dichte verglichen.

3.3.4.1 Globale Clusterstruktur

Als erster Überblick über die Koevolutionssignale innerhalb des Ribosoms wurden die einzelnen Z-Score-Matrizen in Evolutionsmatrizen übersetzt (vergl. Kapitel 3.2.6). Das Ergebnis der Cluster-Analyse ist in Abbildung 13 gezeigt.

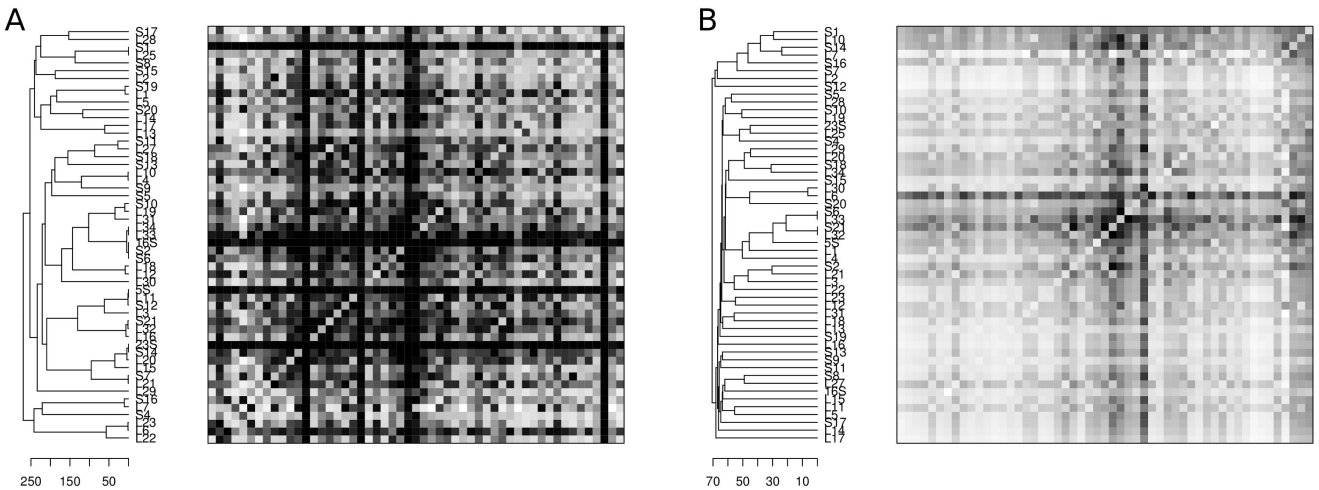

Abbildung 13: Die kolorierten Graphen zeigen die globale Übersicht aus Evolutionsmatrizen für den Vergleich aller ribosomalen Proteinpaarung (X-/Y-Achse) entsprechend dem jeweiligen Dendrogramm sortiert. In (A) ist die Evolutionsmatrix aus dem jeweiligen Maximum der intermolekularen Z-Score-Matrizen ermittelt. Die Evolutionsmatrix aus dem mittleren Z-Score-Signal ist in (B) gezeigt. Links neben der Evolutionsmatrix ist das jeweilige Ergebnis der Cluster-Analyse als Dendrogramm gezeigt.

Die Betrachtung der Dendrogramme zeigt zunächst, dass sich die Cluster aus der `hclust`-Berechnung des maximalen Koevolutionssignals deutlich von dem des mittleren Signals unterscheidet. Beiden ist jedoch gemein, dass sich keine Unterteilung in die Moleküle der großen und kleinen Untereinheit des Ribosoms ergibt. Diese Beobachtung zeigt, dass sich die triviale

Annahme der biochemischen Interaktion als stärkster Auslöser für Koevolution nicht bestätigen lässt. Um solch eine globale Betrachtung auf andere Weise zu erklären wurden die Ergebnisse dieser Evolutionsmatrizen mit den Ergebnissen der GNM-Analyse verglichen (Kapitel 5.3.1.2).

3.3.4.2 Lokale Netzwerkstruktur

Um einen Eindruck der lokalen Netzwerkstruktur in den koevolutionären Netzwerken aus den MI-Matrizen zu erhalten wurden die einzelnen Kenngrößen (Kapitel 3.2.5) mit denen von zufällig generierten Netzwerken verglichen. Dabei wurden die Netzwerkmaße, sofern sie für jeden Knoten einzeln berechnet wurden, zunächst über die Knoten gemittelt und anschließend mit den gemittelten der zufälligen Netzwerke verglichen. Eine beispielhafte Analyse für das zufällig ausgewählte ribosomale Protein L3 ist in Abbildung 14 gezeigt.

Bei der Betrachtung der einzelnen Netzwerkmaße fällt auf, dass sich zwischen 80 % und 100 % Dichte keine Maße für das Netzwerk des L3-Proteins bestimmen lassen, die Unterschiede in der Topologie des Protein-Netzwerkes im Vergleich zu zufälligen Netzwerken zeigen. Diese Beobachtung zeigt sich bei fast allen untersuchten Netzwerken [Daten nicht gezeigt]. Da es bei der MI-Berechnung zwischen konservierten Spalten und allen anderen Positionen innerhalb des Proteins zu einer MI von Null kommt, existieren je nach Konservierungsgrad der Sequenz unterschiedlich viele MI-Werte gleich Null. Aus dieser fehlenden Information leitet sich ein Z-Score von Null ab, der bei der Berechnung der Netzwerkmaße dann zu einem Sprung in der Dichte führen kann. Bei Betrachtung der Netzwerkmaße im Vergleich zwischen zufällig generierten und dem realen Netzwerk zeigt sich in vier der berechneten Netzwerkmaße (Betweenness, Clusteranzahl, mittlere Pfadlänge und dem mittleren Grad der Knoten) eine sehr gute Übereinstimmung mit dem mittleren Netzwerkmaß von zufälligen Netzwerken. Es lässt sich demnach keine Dichte ableiten bei der die Topologie der realen Netzwerke von der Topologie zufälliger Netzwerke abweicht. Die Closeness (Abbildung 14 (D)) und Gradverteilung (Abbildung 14 (E)) in realen Netzwerken zeigen jedoch signifikante Unterschiede zu den Maßen zufälliger Netzwerke. Im Gegensatz zu den zufälligen Netzwerken kann man in den Z-Score-Netzwerken erkennen, dass mit steigender Dichte die mittlere Gradverteilung nicht kontinuierlich sondern sprunghaft ansteigt. Bei geringen Dichten sind eher Knoten mit relativ hohem Grad zu beobachten, als in zufälligen Netzwerken. Zusätzlich werden Knoten mit hohem Grad eher höher vernetzt als Knoten mit kleinem Grad. Dieses Verhalten ist aus skalenfreien Netzwerken bekannt, in denen sich die Vernetzungswahrscheinlichkeit mit dem schon vorhandenen Grad eines Knoten erhöht, woraus sich die sogenannte Hub-Bildung ableiten lässt [126]. Um diese Beobachtung zu verifizieren, wurde exemplarisch die Grad-Grad-Korrelation der Netzwerke untersucht. Wie Newman 2002 [101] beschrieben hat ist die Grad-Grad-Korrelation in zufälligen Netzwerken 0, sodass die Grad-Grad-Korrelation als weiteres Kriterium bestimmt wird, um auszuschließen das es sich bei den Netzwerken aus den Z-Score-Matrizen um zufällige Netzwerke handelt. Dazu gehören sowohl komplett zufällige Netzwerke nach Erdös et al. [44], als auch zufällige skalenfreie Netzwerke nach Albert et al. [1]. Newman beschreibt weiter, dass eine Korrelation innerhalb der Grad-Grad-Korrelation (assortativ) vornehmlich in sozialen Netzwerken zu beobachten ist, wohingegen komplexe biologische Netzwerke eher eine negative, also anti-korrelierte Verteilung aufweisen (disassortativ) [101]. Diese Untersuchung wurde hier auf die Z-Score-Netzwerke angewendet. Die Ergebnisse der Analyse sind in Abbildung 15 gezeigt. Da die Z-Score-Matrizen

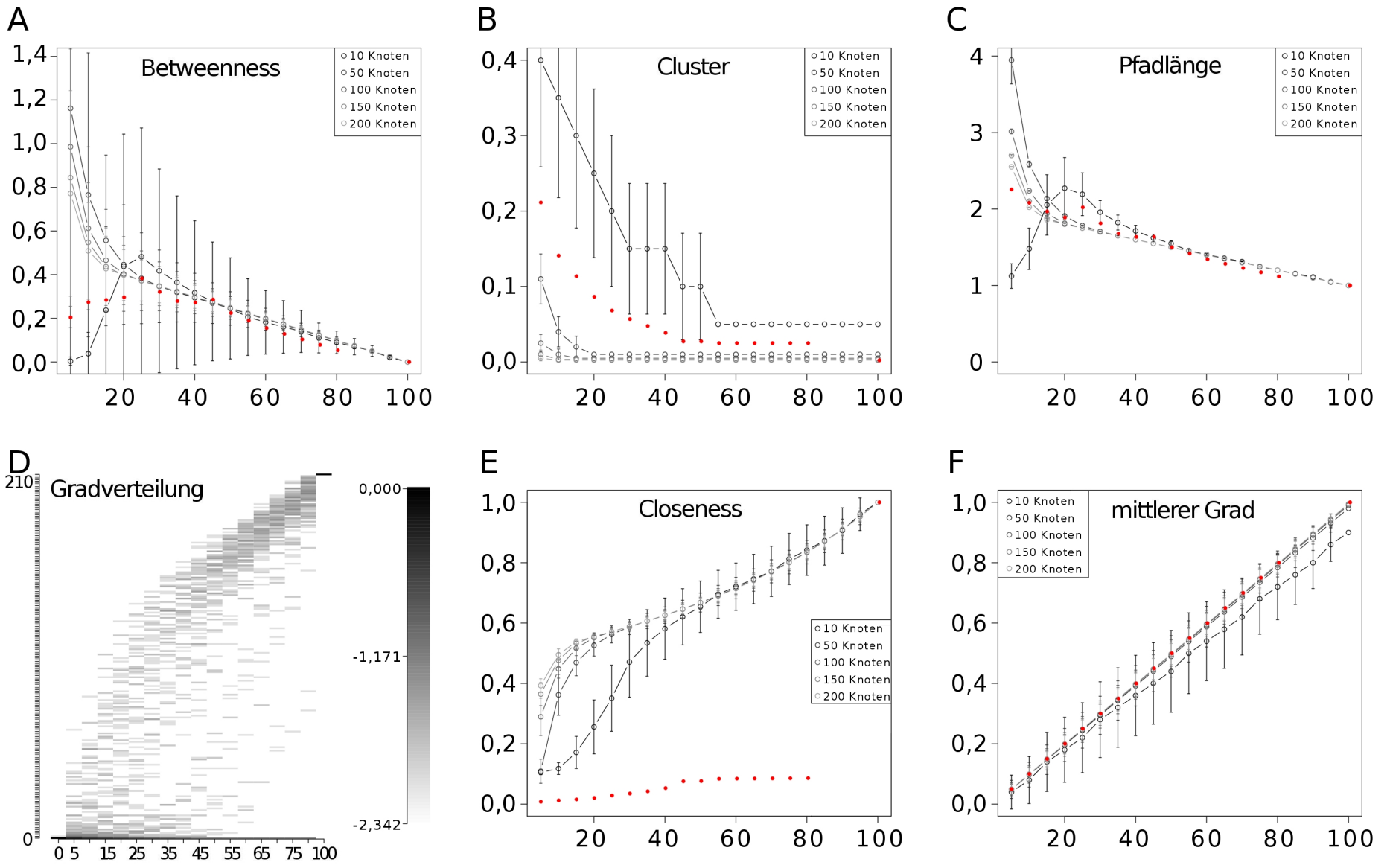

Abbildung 14: Vergleich des Mittelwerts der Netzwerkmaße von zufälligen Netzwerken (grau) mit den Netzwerkmaßen des Z-Score-Netzwerkes des ribosomalen Proteins L3 (rot). Für die Darstellung der Netzwerkmaße wurden diese jeweils durch die Anzahl der Knoten geteilt und über der Dichte in Prozent [%] aufgetragen. In (A) ist die Betweenness, in (B) die normierte Anzahl der Cluster, in (C) die durchschnittliche Pfadlänge, in (D) die Gradverteilung (zur Verdeutlichung wurde der Logarithmus zur Basis 10 aufgetragen), in (E) die Closeness und in (F) der durchschnittliche Grad gezeigt. Die Fehlerbalken geben die Standardabweichung an.

zum Teil sehr viele Nullen enthielten, wurde die Dichte als Kontrollgröße mitberechnet und auch analysiert.

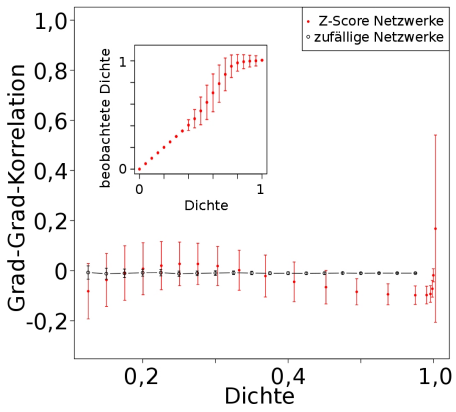

Abbildung 15: Die mittlere Grad-Grad-Korrelation in den Z-Score-Matrizen über der mittleren Dichte der Netzwerke aufgetragen. In rot ist die mittlere Grad-Grad-Korrelation zufälliger Netzwerke gezeigt. Als Inset im Graphen ist die mittlere Dichte der Z-Score-Matrizen gezeigt. Die Fehlerbalken zeigen die Standardabweichung.

Es fällt auf, dass die Dichte in den Graphen der Z-Score Netzwerke teilweise sehr stark von der erwarteten Dichte abweicht. Diese Beobachtung folgt aus der ungleichmäßige Verteilung der berechneten Z-Scores und dem hohen Anteil an Nullen in den Z-Scores. Dadurch werden die Netzwerke mit steigendem Schwellenwert nicht so gleichmäßig vernetzt wie erwartet.

Die Analyse der Grad-Grad-Korrelation zeigt bei Dichten bis ca. 80 % keine signifikante Abweichung von Null, woraus ein nicht von zufälligen Netzwerken verschiedenes Verhalten geschlossen werden kann. Bei Dichten von mehr als 95 % zeigen die Netzwerke disassortatives Verhalten, ausgedrückt durch eine negative Grad-Grad-Korrelation. Johnson et al. beschrieben einen Wert von ca. $-0{,}12$ für Protein-Protein-Netzwerke [73]. Dabei handelt es sich bei den Untersuchungen von Johnson et al. zwar um Protein-Protein-Interaktionen und nicht um die Koevolution einzelner Residuen, da Koevolution jedoch in interagierenden Proteinen durch direkte als auch indirekte Wechselwirkung von Residuen zu beobachten ist [135], wurde dieses Maß verwendet um die Dichte der Z-Score-Netzwerke zu bestimmen bei der biologisch relevante Netzwerkstrukturen zu vermuten sind. Die Grad-Grad-Korrelation von den hier untersuchten Netzwerken liegt bei $\sim$87 % Dichte bei $-0{,}094$, sinkt bis $\sim$95 % Dichte auf $-0{,}097$, dann wird sie wieder größer. In diesem Dichtebereich liegt das disassortative Verhalten der untersuchten Netzwerke also am nächsten an dem von Johnson et al. [73] beschriebenen Wert. Daher wurden die folgenden Untersuchungen auf Netzwerke mit Dichten zwischen 80 - 95 % beschränkt.

3.3.4.3 MI-Netzwerke und Antibiotika-Bindestellen

Die zunächst sehr allgemeinen Analysen (Kapitel 3.3.4.2) sollen nun mit Hilfe von Antibiotika-Bindepositionen im Ribosom detaillierter durchgeführt werden. Ergebnisse von Wilson et al. zeigen einen Bindemechanismus für Makrolid-antibiotika, der über *R. durans* und *M. marismor-*

tui Organismen konserviert vorliegt [138]. Solche konservierten Bindemechanismen können, wie Untersuchungen an der HIV-1 Protease zeigen [81] zu induzierter Koevolution führen. Daher lässt sich vermuten, dass ein solch speziesübergreifender Mechanismus, den Wilson et al. beschreiben, zu koevolutionären Signalen im Ribosom führen können. Diese sollen im Folgenden mit Hilfe der Netzwerkmaße identifiziert werden. Die koevolutionären Signale, die in dieser Arbeit untersucht wurden, stammen alle aus dem 70 S Ribosom von Bakterien. Da sich der Bindemechanismus von Makrolid-Antibiotika nicht nur auf das bakterielle Ribosom bezieht, sondern auch auf Bindung am Ribosom von Archaeen (*M. marismortui*) [138], soll untersucht werden, ob sich auch Muster in den koevolutionären Signalen für Bakterien behandelt mit anderen Antibiotika finden lassen.
Um solche Signaturen in den Alignments der ribosomalen Moleküle zu beobachten, wurden verschiedene PDB-Strukturen, die am Ribosom gebundene Antibiotika enthalten (vergl. Tabelle 2) auf außergewöhnliche Netzwerkmaße hin untersucht. Die Bindepositionen der einzelnen Antibiotika wurden dabei innerhalb des Ribosoms identifiziert, indem ein 13 Å Radius um das Antibiotikum bestimmt wurde. AS und die Nukleotide, die innerhalb dieses Radius lagen wurden als potentielle Interaktionspartner definiert. Die so identifizierten Residuen werden im Folgenden genauer mit Hilfe der Netzwerkmaße innerhalb der Z-Score-Netzwerke untersucht. Dabei wurden die Netzwerke mit einer Dichte zwischen 80 und 95 % untersucht, da hier eine Protein-typische Disassortativität gefunden wurde (Abbildung 15) [73]. Für die detaillierte Analyse wurden dabei jeweils die Maße der intramolekularen Netzwerke analysiert. Dabei wurden Antibiotika untersucht, die entweder an der 30 S oder der 50 S Untereinheit des bakteriellen Ribosoms binden. Die Namen der untersuchten Antibiotika, die bindende Untereinheit und der PDB-Code, aus dem die interagierenden Residuen extrahiert wurden, sind in Tabelle 2 zusammengestellt:

Tabelle 2: Übersicht der untersuchten Antibiotika. Aufgelistet sind jeweils die bindende Untereinheit des Ribosoms, der PDB-Code und der Name des Antibiotikums.

Untereinheit	PDB-Code	Antibiotikum	Organismus
50 S	3OI3	Telithromycin	*T. thermophilus*
	3OHZ	Azithromycin	*T. thermophilus*
	3OH5	Chloramphenicol	*T. thermophilus*
30 S	3DF1	Hygromycin	*E. coli*
	2QAL	Neomycin	*E. coli*
	2QB9	Gentamycin	*E. coli*
	2QUO	Spectinomycin	*E. coli*
	1VS5	Kasugamycin	*E. coli*
	1HNW	Tetracyclin	*T. thermophilus*
	1HNX	Pactamycin	*T. thermophilus*

Von den intramolekularen Netzwerken der mit den in Tabelle 2 aufgelisteten Antibiotika in Kontakt stehenden Moleküle wurden die Netzwerkmaße bestimmt, die sich bei der größten Dichte,

die kleiner als 95 % war, ergab. Eine Übersicht der gefundenen Dichten in den intramolekularen Netzwerken ist in Tabelle 3 zusammengestellt.

Tabelle 3: Dichte der Netzwerke aus den *Z*-Score-Matrizen. Gezeigt ist jeweils die größte Dichte, die kleiner ist als 95 %. Aufgelistet sind alle Moleküle, in denen Residuen innerhalb des 13 Å Interaktionsradius der Antibiotika liegen.

Molekül	16 S	S13	S12	S7	S11	S18	S2	S15
Dichte [%]	25,0	65,6	36,6	50,3	68,8	76,5	75,3	81,0
Molekül	S6	S5	23 S	L27	L32	L4	L22	
Dichte [%]	75,7	61,0	37,7	55,7	81,2	65,3	75,7	

Zunächst ist zu erkennen, dass nur 15 der insgesamt ca. 57 Moleküle des Ribosoms in potentiellem Kontakt mit den hier untersuchten Antibiotika stehen. Dabei finden sich die von Carter et al. [25] beschriebenen Moleküle der A-, P- und E-Site des Ribosoms bis auf drei Ausnahmen wieder. Einzelne ribosomale Proteine wie das S9 konnten nicht untersucht werden, da der `HMMER2.0`-Algorithmus zum Zeitpunkt der Untersuchung nur über den Trainingsdatensatz (vergl. Kapitel 3.2.3) des N-Terminus von S9 verfügte und so keine Sequenzen für den bindenden C-Terminus des S9 Proteins identifiziert werden konnten. Andere Moleküle, die nicht mit der A-, P- oder E-Site in Verbindung gebracht werden können, wie zum Beispiel S5, wurden jedoch bei der Antibiotika-spezifischen Strukturanalyse auch von Carter et al. identifiziert [25]. Auch Pioletti et al. [109] haben ribosomale Komplexe in Verbindung mit Antibiotika-Bindung untersucht und dabei zum Beispiel die Bindung von Tetracyclin an der kleinen ribosomalen Untereinheit über S4, S9 und S17 beschrieben. Keines dieser Proteine kann in unserem Datensatz über die Entfernung von maximal 13 Å identifiziert werden und auch Carter et al. [25] zeigen keines der Proteine in der Beschreibung der A-Site, die durch Tetracyclin geblockt wird. Für Tetracyclin werden von Pioletti et al. [109] bis zu vier Bindestellen beschrieben, von denen in der hier untersuchten Struktur (PDB-Code 1HNW) aber nur zwei besetzt sind. Die große Entfernung der zusätzlich beschriebenen Proteine in Verbindung mit Tetracyclin zu den hier besetzten Bindepositionen könnten ein Hinweis auf transiente Bindepartner darstellen [17]. Daher werden diese Proteine in dieser Arbeit nicht näher analysiert.

Die über die Disassortativität festgelegte Schwelle für die zu untersuchende Dichte der Netzwerke von 80-95 % (vergleiche Abbildung 14) unterschreiten die meisten hier gezeigten Netzwerke aus den *Z*-Score-Matrizen. Diese Beobachtung hängt mit der Eigenschaft der Netzwerke zusammen, dass die Dichte durch den hohen Anteil an sehr niedrigen Werten sprunghaft ansteigen kann, sodass zum Beispiel das Netzwerk der 16 S rRNA ab einer Dichte von 25 % direkt auf eine Dichte von über 95 % ansteigt [Daten nicht gezeigt]. Aus diesem Grund werden nur die Netzwerke von S15 und L32 detailliert untersucht. Auffällig ist zunächst, dass sich das Protein L32 bei allen hier untersuchten Antibiotika als Bindepartner findet. Bei näherer Betrachtung finden sich für alle drei Antibiotika, die an der großen Untereinheit binden, sehr ähnliche Bindepartner, nämlich L3, L4 und L32, wobei für L3 kein Alignment zur Verfügung steht und die Dichte des L4-Netzwerks zu gering ist, weshalb die detaillierte Analyse auf L32 beschränkt wird. Die große Ähnlichkeit der Bindeinteraktionen der drei untersuchten Antibiotika in der großen

Untereinheit ist von Wilson et al. [138] gezeigt worden und impliziert ein mögliches Muster in der Koevolution der so identifizierten Positionen in den Molekülen. Die Abbildung 16 zeigt die Netzwerkmaße innerhalb des L32-Netzwerkes an den spezifischen Positionen. Dabei sind Positionen, die alle drei Antibiotika binden, hervorgehoben.

Für die intramolekularen Netzwerke ließen im Laufe dieser Arbeit keine Maße finden, die an den Knoten der Antibiotika-Bindestellen eine Korrelation zwischen Maß und Position zeigen. Somit kann eine potentielle Antibiotika-Bindestelle allein über das Netzwerkmaß der koevolutionären Signale nicht identifiziert werden.

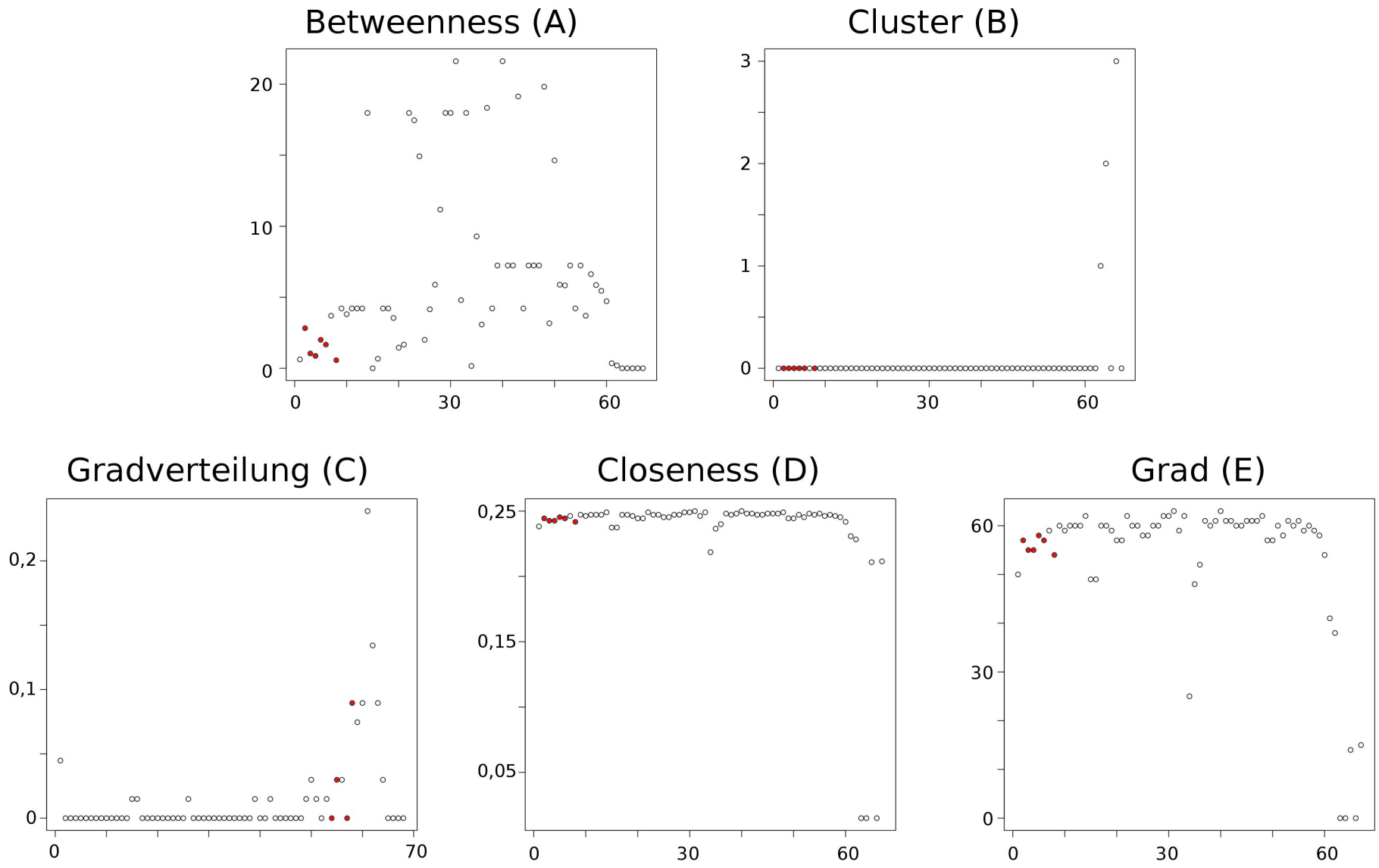

Abbildung 16: Netzwerkmaße des ribosomalen Proteins L32 für jedes Residuum (X-Achse). In rot sind die Antibiotika bindenden Residuen markiert.

3.4 Diskussion

3.4.1 Einflüsse auf die MI-Berechnung

Für die Berechnung der MI sind vier sehr einflussreiche Effekte zu beachten, auf die in dieser Arbeit im Einzelnen eingegangen wurde. Dabei handelt es sich zunächst um den Einfluss des Alignments. Dieser konnte für die rRNA-Alignments durch den Vergleich der daraus resultierenden phylogenetischen Bäume mit bereits publizierten Bäumen von Wu et al. [143] überprüft werden. Für die Protein-Alignments bestehen keine äquivalenten Vergleichsmöglichkeiten, sodass die bei den RNA-Alignments verwendeten Standardparameter auch in diesen Alignments verwendet wurden.

Um den Einfluss der Stichprobengröße und des phylogenetischen Effektes zu reduzieren, wurden verschiedene Methoden eingeführt (vergleiche Kapitel 3.2.2). Im Rahmen dieser Arbeit wurde ein einfaches Nullmodell für die Evolution des bakteriellen Ribosoms etabliert (Kapitel 2, [136]), das über die Bildung von Z-Scores eine durchschnittliche zufällige MI auf Grundlage der verwendeten Daten abschätzen kann, wodurch diese beiden oben genannten Effekte durch implizite Berechnung im Nullmodell berücksichtigt werden.

Den vierten Einfluss auf die berechneten Ergebnisse der MI-Analyse stellen die Gaps aus den Alignments dar. Dabei wurde dieses Problem bisher hauptsächlich ignoriert, da die Interpretation der Gaps im Rahmen der MI-Analyse offen ist. Das biologische Phänomen, das einen Verlust oder einen Gewinn an Informationen innerhalb biologischer Sequenzen darstellt, kann in der Alignment-basierten Informationstheorie bisher nicht abgebildet werden.

Die hier vorgestellten Verfahren versuchen das Gap-Problem dadurch zu umgehen, nicht die Interpretation der Gaps, sondern der verbleibenden Symbole vorzunehmen. Diese Verfahren sollen den Informationsgewinn aus den verbleibenden oder den hinzugewonnenen Symbolen bewerten, statt den Verlust (Deletion) oder Gewinn (Insertion) an Information gesamt. Die Methoden erfüllen dabei diese Aufgabe abhängig von der beobachteten Anzahl an Gaps unterschiedlich gut. Bei hohem Gap-Anteil wird die Interpretation der ESMI und DEMI schwierig, da hier die theoretischen Grenzen der MI verschoben werden. Die DEMI wird in diesen Fällen schnell negativ und die ESMI erzeugt Werte, die weit über dem theoretischen Maximum der MI liegen können. Die SUMI ist eine ähnliche Methode wie die ORMI, sie betrachtet das Gap-Symbol jedoch nicht als zusätzliches Symbol des Alphabets, sondern berechnet aus dem verbliebenen Subset die resultierende MI. Eine detaillierte Klärung konnte auch im Rahmen dieser Arbeit nicht gezeigt werden. Es ist aber ein erster Schritt hin zu einer Interpretation etabliert worden.

3.4.2 Sensitivität intermolekularer MI

Die Berechnung intermolekularer MI hat zusätzlich zu den schon erwähnten Einflüssen eine weitere Einschränkung. Für das Erstellen der intermolekularen Berechnungsgrundlage müssen die intramolekularen Alignments weiter prozessiert werden. Die hier vorgestellte Methode verstärkt durch die Reduktion und Kombination der Sequenzen den phylogenetischen Effekt. Diese Verstärkung des Effektes wird durch das hier verwendete Nullmodell (Kapitel 2) mit in die Bestimmung der Signifikanz der Ergebnisse einbezogen, sodass sich mit der Verwendung der

Z-Scores für die Betrachtung der MI-Ergebnisse immer eine Korrelation ergibt, die zumindest in der Tendenz stimmt, da sich in jedem Fall der Effekt mit einem Pearson-Koeffizienten von ca. 0,5 (vergleiche Abbildung 10) berechnen lässt. Sollte eine differenzierte Analyse auf Grundlage der MI-Berechnung durchgeführt werden, müsste dieser Effekt durch einen Korrekturterm berücksichtigt werden. Da dieser Effekt für die tendenzielle Analyse in dieser Arbeit jedoch als marginal angenommen wurde, ist ein Korrekturterm hier nicht notwendig.

3.4.3 MI-Betrachtung

Für die detaillierte Analyse der Z-Scores aus den MI-Berechnungen ist eine Betrachtung des Einflusses des kombinatorischen Ansatzes für die intermolekulare MI nötig, sowie eine differenzierte Analyse der Alignmentpositionen mit hohem Gap-Anteil. Außerdem werden die in der Pfam [47, 125] verfügbaren HMM mit immer größeren Datensätzen trainiert, sodass eine erneute Identifikation der ribosomalen Proteine verbesserte Ergebnisse und größere Datensätze als Berechnungsgrundlage ermöglichen würde. Da zum Zeitpunkt der Identifikation der ribosomalen Proteine innerhalb der Genomsequenzen die Trainingsdatensätze der HMMs zum Teil nur aus vier Sequenzen bestanden, können die Alignments dieser Proteine nicht für Einzelanalysen der MI herangezogen werden, ohne ein verfälschtes Ergebnis zu liefern.

Diese Arbeit hat gezeigt, dass in der Tat Antibiotika-Bindestellen nicht aus MI-Berechnungen extrahiert werden können, wie es die Ergebnisse von Wilson et al. [138] durch den gemeinsamen Bindemechanismus implizieren könnten. Dies ist nicht überraschend, da die überwiegende Mehrheit der Organismen unseres Datensatzes keinem Selektionsdruck durch Antibiotika unterlag. Daher können die verschiedenen Interaktionen der AS untereinander und die darauf beruhenden Koevolutionsmuster die durch Antibiotika-Bindung induzierte Koevolution überdecken. Außerdem ist es nicht möglich, den untersuchten Datensatz der ribosomalen Molekülsequenzen getrennt nach gebundenem und ungebundenem Antibiotikum zu betrachten, wie es bei der HIV-1 Protease möglich war. Dadurch kann ein durch Antibiotika induzierter Effekt nicht durch die Bildung eines Differenzsignals sichtbar gemacht werden. Die Vielzahl von Signalen der MI sollte daher durch das Bilden von Netzwerken auf Strukturen höher hierarchischer Mechanismen untersucht werden. Gleichzeitig haben die „typischen“ Bindestellen der Antibiotika aber auch in ihren durch AS-Interaktion vermittelten MI-Netzwerken kein differenziertes Signal für unterschiedliche Antibiotika gezeigt. Daher können wir leider aus den MI-Netzwerk-Ergebnissen nicht ablesen welche potentielle Resistenz-Evolution durch MI-Netzwerk-Cliquen eingeschränkt und somit unterdrückt wäre. Im Fall der Antibiotika gegen das bakterielle Ribosom ist daher anders als bei der HIV-1 Protease [15] und reversen Transkriptase [62] leider keine weitere Vorhersage über „lohnende“ Targets möglich.

3.4.4 MI-Netzwerke

Aus den Ergebnissen der MI-Analyse wurde durch Bildung von Evolutionsmatrizen eine globale Betrachtung abgeleitet, die jeweils das maximale oder durchschnittliche Evolutionssignal der verglichenen ribosomalen Moleküle beinhaltet. Die dadurch entstehenden Cluster sind nicht durch triviale Erklärungsansätze zu erfassen, da es hier zu unterschiedlichen Ergebnissen in den Clustern kommt (vergleiche Abbildung 13 (A) und (B)). Beide zeigen jedoch, dass die Annahme, dass Koevolution hauptsächlich durch räumliche Nähe induziert ist, widerlegt werden kann.

Auch zeigt sich in keiner der beiden Betrachtungen die Unterteilung in Moleküle der kleinen und großen Untereinheit. Diese Ergebnisse deuten eher darauf hin, dass sich Koevolution in makromolekularen Komplexen durch langreichweitige Interaktionen, die durch andere Mechanismen induziert werden, beobachten lässt.

Um solche möglichen langreichweitigen Effekte genauer aufzuklären, wurden makromolekulare Dynamik und thermodynamische Stabilität untersucht und mit den Koevolutions-Ergebnissen abgeglichen (Kapitel 5.3.2 und Kapitel 5.3.1.1).

3.4.5 Lokale Netzwerke

Da es sich bei den Z-Score-Matrizen um vollständig besetzte Matrizen handelt, ist es schwer einen nicht vollständig vernetzten Graphen zu erstellen. Dem wurde hier damit entgegnet, dass aus jeder Matrix Netzwerke verschiedener Dichten erzeugt wurden. Dabei sollten anschließend biologisch relevante Netzwerke durch den Vergleich mit Netzwerkmaßen aus zufälligen Netzwerken identifiziert werden. Hier zeigte sich eine erwartet große Übereinstimmung mit zufälligen Netzwerken. Dabei sind die am deutlichsten abweichenden Maße die Closeness und die Gradverteilung, die den linearen Anstieg des Mittels einer Gaußschen Verteilung um den mittleren Grad des Netzwerkes nicht zeigte. Diese Beobachtungen führten zur Verwendung der Grad-Grad-Korrelation als Identifikation der Dichte.

Auch in diesen Netzwerken ging es um die Identifikation biologisch relevanter Strukturen, die für das Ribosom untersucht sind. Dabei wurde auch in den Netzwerken bei den spezifischen Dichten untersucht, ob sich an AB-Bindestellen relevante Maße ergeben. Weder in den intramolekularen noch in den intermolekularen Netzwerken konnten Netzwerkmaße identifiziert werden, die ein charakteristisches Verhalten an genau diesen Positionen zeigen.

Die Identifikation von AB-Bindestellen oder dem koevolutionären Einfluss dieser Positionen ist mit den hier durchgeführten Analysen zunächst nicht möglich. Die Interpretation globaler Maße, wie Evolutionsmatrizen im Vergleich mit GNM-Analysen, kann schon publizierte Ergebnisse zeigen [62]. Daraus ist zu schließen, dass die mit Hilfe der HMM identifizierten Sequenzen die Möglichkeit bieten allgemeine Tendenzen in makromolekularen Komplexen zu finden. Die Identifikation einzelner Residuen oder Domänen für Medikamenten-Design oder die funktionelle Aufklärung der Struktur setzt jedoch eine differenzierte Beobachtung von Koevolution innerhalb Sequenzen mit und ohne der zu untersuchenden Eigenschaft voraus (vergleiche Kapitel 4.3).

Zu den Einzelnetzwerken ist im Bereich der intermolekularen MI zu erwähnen, dass es sich hier im Detail betrachtet eigentlich um bipartite Netzwerke handelt, die durch ihren unterschiedlichen Aufbau nicht mit den konventionellen Netzwerkmaßen einfacher Netzwerke zu vergleichen sind. In der MI-Bestimmung ergibt sich dadurch jedoch ein weitaus schwierigeres Problem, denn durch das Ausblenden der intramolekularen MI bei Betrachtung der bipartiten Netzwerke kann es hier zu Fehlinterpretationen kommen. Eindeutige Signale die zum Beispiel für zwei Residuen im intermolekularen Vergleich bestimmt werden, müssen nicht zwangsläufig auch für einen intermolekularen Zusammenhang stehen. Vielmehr kann es sein, dass ein intramolekulares Signal zu einer Identifikation von intermolekularen Signalen führt. Ein ähnliches Phänomen kann man bei der Identifizierung im intramolekularen Bereich beobachten.

Wenn es um die Koevolution zweier weit entfernter Residuen geht, könnte diese Koevolution auch durch eine Verkettung näher gelegener Signale entstanden sein. Hierzu gibt es Ansätze zur Unterscheidung von direkter und indirekter MI von Burger et al. [21].

4 Biochemische MI in der HIV-1 Protease

4.1 Einleitung

Wie im Kapitel 3.1 schon erwähnt, wird die MI als Maß der Koevolution verwendet. Dabei wurden in biologischen Anwendungen der MI bisher hauptsächlich Primärsequenzen von Proteinen, Desoxyribonukleinsäuren (DNA) oder Ribonukleinsäuren (RNA) analysiert. Neue Studien von Gao et al. [51] zeigen, dass auch eine umkodierte Sequenz Einblicke in die Struktur von Proteinen ermöglichen kann. Dazu werden zum Beispiel die Aminosäuren in die biochemischen oder physikalischen Eigenschaften umkodiert. In dieser Arbeit wurde die Primärsequenz eines aus dem Humanen Immunodefizienz-Virus (HIV) stammenden Proteins untersucht. Das HIV ist der Auslöser von AIDS (*acquired immunodeficiency syndrome*) [92]. Das Virus kann durch seine hohe Mutationsrate [76, 89] sehr schnell Resistenzen generieren und entgeht so vielen Therapieansätzen. Eines der intensivst erforschten Proteine des HIV ist die HIV-1 Protease, die zur Gruppe der Aspartat-Proteasen gehört und durch die Spaltung von Proteinvorläufern diese in ihre funktionelle Form überführt. Wird diese Protease inhibiert, kann die Vermehrung des Virus effizient unterdrückt werden [94, 6]. Damit stellt die Protease ein ideales Ziel für die Bekämpfung von HIV mittels Proteaseinhibitoren (PI) dar. Durch diese Eigenschaft wurde auch die strukturelle Aufklärung der HIV-1 Protease motiviert und unter anderen von Wlodawer et al. [139] detailliert gezeigt.

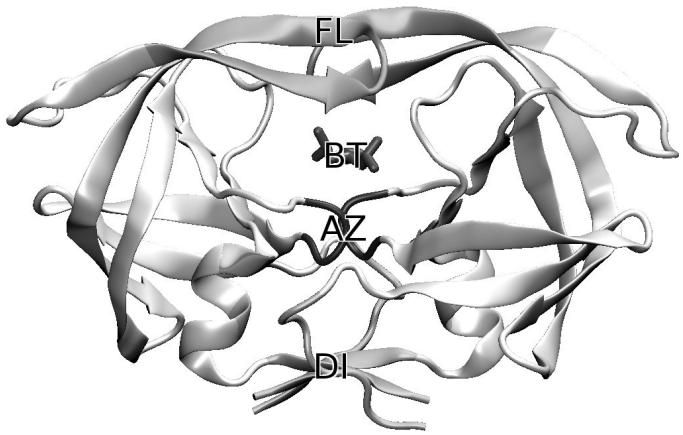

Abbildung 17: Struktur der HIV-1 Protease. Einige wichtige strukturelle Merkmale sind grau hervorgehoben: Die Flap-Region (FL), die Bindetasche (BT) ist zusätzlich durch 2-Amino-4-Methyl-Pentan-1-Ol markiert, das aktive Zentrum (AZ) und das Dimerisierungsinterface (DI).

Die wichtigsten Strukturmerkmale der HIV-1 Protease sind in der Abbildung 17 am Beispiel der Struktur von Mahalingam et al. [86] mit dem PDB-Code 1DAZ gezeigt, in der ein kleiner Ligand (2-Amino-4-methyl-pentan-1-ol) in der Bindetasche des Enzyms zu sehen ist. Für die enzymatische Aktivität der HIV-1 Protease ist die Flexibilität der *Flaps* sehr wichtig [62, 80, 72], die somit ein strukturell wichtigen Teil der Protease darstellen. Bei der HIV-1 Protease han-

delt es sich um ein Homodimer, das über das Dimerisierungsinterface und Wechselwirkungen zwischen dem anschließenden Turn und der gegenüberliegenden α-Helix seine Quartärstruktur ausbildet [139].

Die strukturelle Aufklärung, auch in Verbindung mit PI, hat über die Identifikation der bindenden Residuen zur Entwicklung von neuen Medikamenten geführt [94, 10]. Außerdem hat die medizinische Erfassung von Patientendaten die Erforschung der HIV-1 Protease erleichtert. Im Besonderen hat hier die Sequenzierung des HIV aus Blutproben von Patienten zu großen Datenbanken geführt [121, 122, 105], die auch bioinformatische Analysen auf Basis der Primärsequenz ermöglichen. Der Zusammenhang zwischen Resistenzen gegen einzelne PI und Mutationen in der HIV-1 Protease Sequenz wurde in verschiedenen medizinischen Studien gezeigt [35, 36, 90, 38]. Auf Grundlage von medizinischen Studien und Datenbanken ist eine immer genauere Identifikation, der Resistenzmutationen möglich [74]. Auf Grundlage dieser Datenbestände konnten Liu et al. zeigen, dass in den Koevolutionsmustern der MI zwei Klassen von Residuen zu finden sind, nämlich einem phylogenetisch variables Cluster (*phylogenetic variance cluster* (PhVC)) und ein Medikamentenresistenz-Cluster (*drug resistance cluster* (DRC)), was die Ergebnisse der manuellen Analysen der vorangegangenen Untersuchungen bestätigt [81]. Diese Ergebnisse sollen anhand der Umkodierung der Primärsequenzen von HIV-1 Proteasen in einen Biochemischen Kontext gesetzt werden.

4.2 Methoden

4.2.1 Biochemische MI (cheMI)

Um den Ursprung des koevolutionären Signals der MI detaillierter zu klären, wurde die MI für jede biochemische Eigenschaft einzeln berechnet. Für diese Berechnung standen zwei unterschiedliche Datensätze für Sequenzen der HIV-1 Protease zur Verfügung.

Die MI (Kapitel 2.2.2) berechnet ein koevolutionäres Signal zwischen Residuen eines Moleküls. Da biochemische Interaktionen zwischen Residuen solch eine Koevolution bedingen können, soll dieser Zusammenhang genauer aufgeklärt werden. Im Rahmen dieser Arbeit wurde eine neue Methode etabliert, die darin besteht, das klassische AS-Alphabet in ein biochemisches zu übersetzen, und durch die Analyse der so neu kodierten Sequenzen mit Hilfe der MI den biochemischen Hintergrund der Koevolution aufzuklären. Die Übersetzung des AS-Alphabets wird durch die Klassifizierung nach Taylor [129], erweitert um die Gruppe `prolin` (siehe Tabelle 4), mit zwei unterschiedlichen Methoden durchgeführt.

1. 1Bit-Translation: Die Sequenzen werden für jede biochemische Eigenschaft binär kodiert. Jede AS wird dabei mit einer 1 übersetzt, wenn sie über die entsprechende biochemische Eigenschaft verfügt, sonst mit einer 0.

2. 8Bit-Translation: Die AS der Sequenzen werden für alle ihre biochemischen Eigenschaften gleichzeitig kodiert (siehe Tabelle 4 ohne `klein` und `geladen`), sodass jede AS in einen binären 8bit-Vektor übersetzt wird, der die Zugehörigkeit zu jeder der biochemischen Eigenschaften enthält.

Aus der ersten Methode entsteht für jede biochemische Eigenschaft ein umkodiertes Alignment, wohingegen die zweite Methode ein einziges Alignment für alle biochemischen Eigenschaften erzeugt. Um die Vollständigkeit der Sequenzen zu erhalten, wurden die Platzhalter-AS, wie *B* und *Z*, die jeweils für ASX beziehungsweise GLX kodieren, nur dann mit einer 1 übersetzt, wenn beide AS ein einheitliches biochemisches Signal zeigten. Der allgemeine Platzhalter *X*, der für eine unbestimmte AS steht, wurde in allen Alignments mit 0 übersetzt.

Tabelle 4: Die biochemischen Eigenschaften und die dazugehörigen AS sortiert nach Gruppengröße. Die hier gezeigte Klassifikation nach Taylor [129] ist erweitert um die Gruppe `prolin`.

Gruppe	Aminosäuren
`hydrophob`	CAGILVKMTHFYW
`polar`	QWYHRKTDECSN
`klein`	CVTDNGASP
`geladen`	HREKD
`aromatisch`	FWYH
`winzig`	GACS
`positiv`	HKR
`aliphatisch`	ILV
`negativ`	ED
`prolin`	P

4.2.1.1 Die 8Bit-Kodierung

In der 8Bit-Kodierung werden nur acht der in Tabelle 4 genannten Eigenschaften verwendet, nämlich: `hydrophob`, `polar`, `aromatisch`, `winzig`, `positiv`, `aliphatisch`, `negativ` und `prolin`.

Dabei wird jeder AS ein Vektor aus Nullen und Einsen zugewiesen, der jeweils das Vorhandensein der acht biochemischen Eigenschaften anzeigt. Als Beispiel für die 8Bit-Kodierung betrachten wir die AS Alanin, die durch den 8Bit-Code `10010000` repräsentiert wird, da sie `hydrophob`, nicht `polar`, nicht `aromatisch`, `winzig`, nicht `positiv`, nicht `aliphatisch`, nicht `negativ` und nicht `prolin` ist.

Die so kodierten AS bilden dann 14 Gruppen mit unterschiedlichen Kombinationen aus biochemischen Eigenschaften. Diese Gruppen wurden römisch nummeriert und sind in folgender Tabelle zusammengefasst:

Tabelle 5: AS-Komposition der 14 biochemischen Gruppen, die durch die 8Bit-Kodierung entstandenen sind.

0	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
X	P	(NQ)	S	R	(ED)	M	(AG)	T	C	K	F	(WY)	H	(LIV)

4.2.2 HIV-Sequenzen

Für die Analyse der HIV-1 Protease standen zum Einen die Sequenzen zur Verfügung, die von der Gruppe um Lee [105, 27] gesammelt und verfügbar gemacht wurden (Lee-DS) und zum Anderen der Datensatz aus der *HIV drug resistance database* (HIV-DS) [115, 116, 122, 121]. Jede der beiden Datenbanken verfügen über ca. 45 000 Sequenzen, wodurch eine gute Statistik gewährleistet ist. Im HIV-DS existieren einige trunkierte oder elongierte Sequenzen, die aussortiert wurden. Nach diesem Aussortieren standen 31 514 Sequenzen von unbehandelten und 13 140 Sequenzen von behandelten Patienten aus der HIV-DS zur Verfügung. Den behandelten Patienten wurden Protease-Inhibitoren (PI) verabreicht, entweder einzeln oder als Kombination verschiedener PI, und die Sequenzen der HI-Viren 3-27 Wochen nach Beginn der Therapie aus Blutproben sequenziert [90, 36, 100, 106, 60]. Der Lee-DS besteht aus 45 160 Sequenzen, die in den Jahren zwischen 1999 und 2002 aus dem Blutplasma von Patienten isoliert wurden, deren Behandlungsstatus nicht genauer bekannt ist. Da in beiden Datensätzen nur Sequenzen mit einer Länge von 99 AS enthalten waren, musste kein Alignment durchgeführt werden.

4.2.3 Hauptkomponentenanalyse

Die Hauptkomponenten-Analyse (*principal component analysis* (PCA)) ist ein mathematisches Verfahren, das auf Pearson (1901) zurückgeht. Dieses Verfahren wird dazu verwendet mehrdimensionale Datensätze niedrig-dimensional darzustellen. Dabei werden durch orthogonale Transformation die Daten so angeordnet, dass die erste Transformation einen „neuen“ Datensatz mit der größtmöglichen Varianz ergibt. Dieser neue Datensatz bildet so die erste Hauptkomponente. Durch weitere orthogonale Transformation werden weitere Datensätze erzeugt, die absteigend ihrer Varianz geordnet werden und so die einzelnen Hauptkomponenten des ursprünglichen Datensatzes bilden. Beschreiben wenige Hauptkomponenten den Großteil der Varianz des ursprünglichen Datensatzes so kann eine Reduktion auf diese Hauptkomponenten durchgeführt werden, um den anfänglichen Datensatz ausreichend zu beschreiben. Die Varianz kann also als Informationsgehalt der Dimension interpretiert werden, den die jeweilige Hauptkomponente abbildet.

Ist keine eindeutige Dominanz eines einzelnen Eigenvektors zu erkennen, die Darstellung der Matrix jedoch in Form eines Vektors gewünscht, kann die Summe der über Ihre Eigenwerte gewichteten Eigenvektoren dazu verwendet werden, ein beliebig genaue Abbildung der Matrix zu erreichen. Für die Darstellung einer Matrix M in einem Vektor $\vec{v}_{\text{rek}}$ kann die Matrix mit Hilfe der Singulärwertzerlegung (*singular value decomposition* (SVD)) [55] in ihre Eigenwerte λ und Eigenvektoren $\vec{v}$ zerlegt werden. Die Rekonstitution der quadratischen, symmetrischen Ausgangsmatrix M erfolgt nach:

$$M = \sum_{i=1}^{N} (\vec{v}_i \cdot \lambda_i \cdot \vec{v}_i^{\,T}) \tag{17}$$

Dabei ist N die Gesamtzahl der Eigenwerte und gleichzeitig die Matrixdimension. Für die Darstellung von M durch einen Vektor $\vec{v}_{\text{rek}}$, sollten mindestens 80 % der Matrix repräsentiert wer-

den. Dafür muss eine obere Grenze I für die Summe der absteigend nach Größe sortierten Eigenwerte λ_i gefunden werden, sodass folgende Bedingung erfüllt ist:

$$0.8 \leq \frac{\sum_{i=1}^{I} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \tag{18}$$

Die anschließende Berechnung des Vektors $\vec{v}_{\text{rek}}$ erfolgt nach:

$$\vec{v}_{\text{rek}} = \sum_{i=1}^{I} \left(\sqrt{\lambda_i} \cdot \vec{v}_i \right) \tag{19}$$

In dieser Arbeit wurde die PCA und die SVD mit Hilfe der Statistik-Software R [113] durchgeführt.

4.3 Resultate

Im Rahmen dieser Arbeit wurde auch die Fähigkeit der MI genutzt, Korrelation in allgemeinen Symbolsequenzen zu identifizieren. Die im Folgenden vorgestellten Ergebnisse beruhen auf einer biochemischen Kodierung der AS-Eigenschaften in dem Alignment der HIV-1 Protease. Dazu wurden die AS wie in Kapitel 4.2.1 beschrieben kodiert, um so mit den veränderten Symbolen der Alignments die biochemische Koevolution innerhalb der HIV-1 Protease zu beschreiben.

4.3.1 1Bit-cheMI

Um das Signal jeder einzelnen biochemischen Eigenschaft in der HIV-1 Protease zu untersuchen, wurde die Sequenzentropie mit der Entropie der jeweiligen biochemischen Eigenschaften verglichen. Die Scatter-Plots aus Abbildung 18 sind durch die jeweiligen Pearson-Koeffizienten in Tabelle 6 zusammengefasst. Dabei fällt die erwartete Korrelation zwischen Entropie der AS-Sequenz und biochemischer Eigenschaft je nach Eigenschaft unterschiedlich stark aus.

In Abbildung 18 zeigt sich eine positive Korrelation zwischen der Entropie der biochemischen Eigenschaften und Sequenzentropie in einem Bereich von ca. 0,2 - 0,6 des Pearson-Koeffizienten. Dass die Korrelation zwischen der Sequenzentropie und der biochemischen Entropie nicht mit der Gruppengröße der biochemischen Eigenschaft zusammenhängt, zeigt der ermittelte Pearson-Koeffizient zwischen der Gruppengröße (in Anzahl AS) und dem jeweils gefundenen Korrelationskoeffizienten der jeweiligen Gruppe von 0,597. Diese Beobachtung lässt den Schluss zu, dass sich hier Signale finden, die durch die Betrachtung des AS-Alphabets allein nicht zu erklären sind.

Tabelle 6: Pearson-Koeffizient zwischen der Sequenzentropie und der Entropie der biochemischen Eigenschaft sortiert nach Größe der Korrelation. Die Anzahl an AS mit der biochemischen Eigenschaft ist unter dem Pearson-Koeffizient eingetragen.

Eigenschaft	klein	aliphatisch	hydrophob	polar	winzig
Pearson	0,598	0,598	0,578	0,457	0,448
Anzahl [AS]	9	3	13	12	4
Eigenschaft	geladen	prolin	neg. geladen	pos. geladen	aromatisch
Pearson	0,363	0,329	0,284	0,239	0,21
Anzahl [AS]	5	1	2	3	4

Es gibt biochemische Eigenschaften, die durch andere vollständig beschrieben werden, zum Beispiel wird die Eigenschaften `geladen`, durch `positiv` und `negativ` beschrieben. Diese Beobachtung führte zu der Frage, ob sich die Eigenschaften reduzieren lassen, ohne dabei Information zu verlieren. Um die Signaländerung durch die Reduktion der Eigenschaften zu analysieren, wurden die ursprünglich neun Eigenschaften nach Taylor [129] auf zwei unterschiedliche Arten reduziert. Die erste Reduktion bestand darin, Gruppen zu entfernen, die vollständig in einer anderen Gruppe enthalten sind. So ist zum Beispiel die Gruppe `winzig` vollständig in `klein` enthalten und wird verworfen. In der zweiten Reduktion wurde jeweils die andere Gruppe verworfen, im Beispiel von `winzig` und `klein` wurde im zweiten Ansatz `klein` verworfen. Im ersten

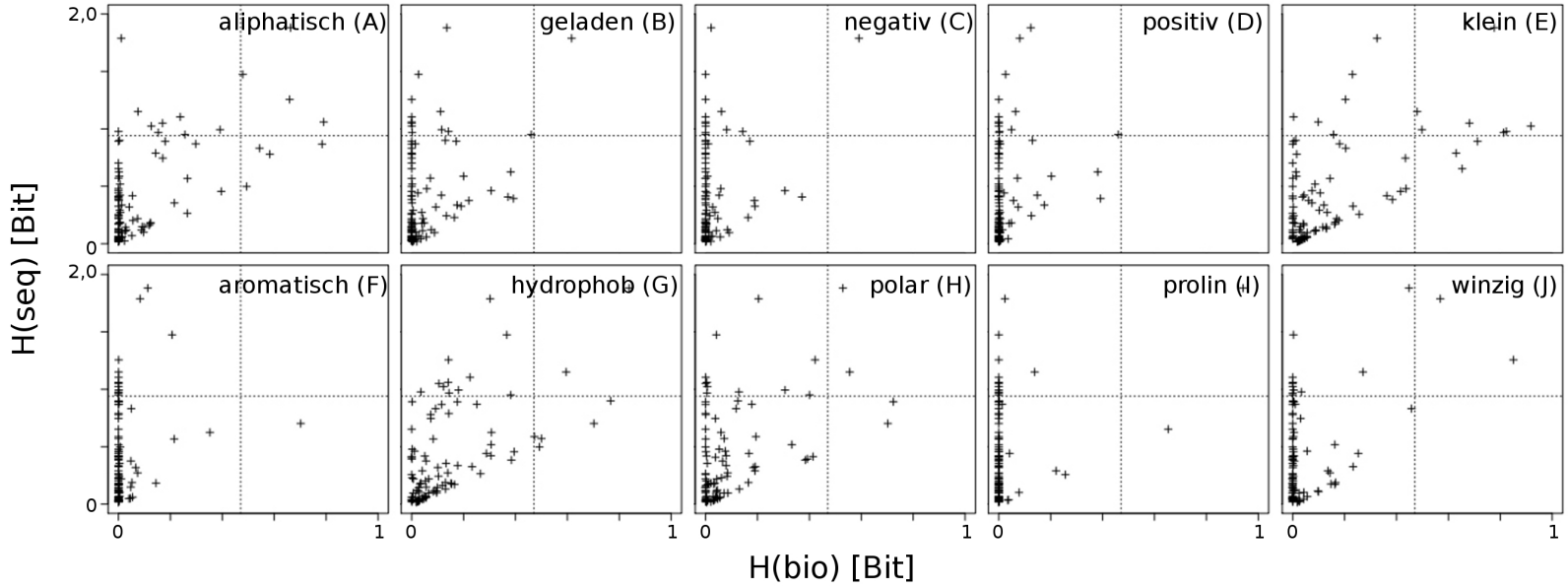

Abbildung 18: Sequenzentropie jedes biochemisch übersetzten Sequenz-Sets (H(bio)) aufgetragen gegen die Sequenzentropie der HIV-1 Protease (H(seq)). Zur Orientierung, ist die halbmaximale aller Werte als gestrichelte Linie gekennzeichnet.

Ansatz sind möglichst überlappende Gruppen erwünscht, wohingegen im zweiten Ansatz möglichst disjunkte Gruppen behalten werden.

Um den Informationsverlust der beiden reduzierten Ansätze zu bestimmen, wurde jeweils eine PCA (Kapitel 4.2.3) für die 1Bit-cheMI Ergebnisse aller biochemischen Eigenschaften durchgeführt und dann mit der PCA der beiden reduzierten Ergebnisse verglichen. Die Signaländerung der Reduktionen wurde über den Vergleich der Varianz der Hauptkomponenten bezüglich des nicht reduzierten Datensatzes ermittelt. Die Gegenüberstellung der drei Gruppen-Sets (alle Eigenschaften nach Taylor Abbildung 19 (A), überlappende Eigenschaften Abbildung 19 (B) und disjunkte Eigenschaften Abbildung 19 (C) sind nach Größe der Varianzen der Hauptkomponenten sortiert in Abbildung 19 gezeigt.

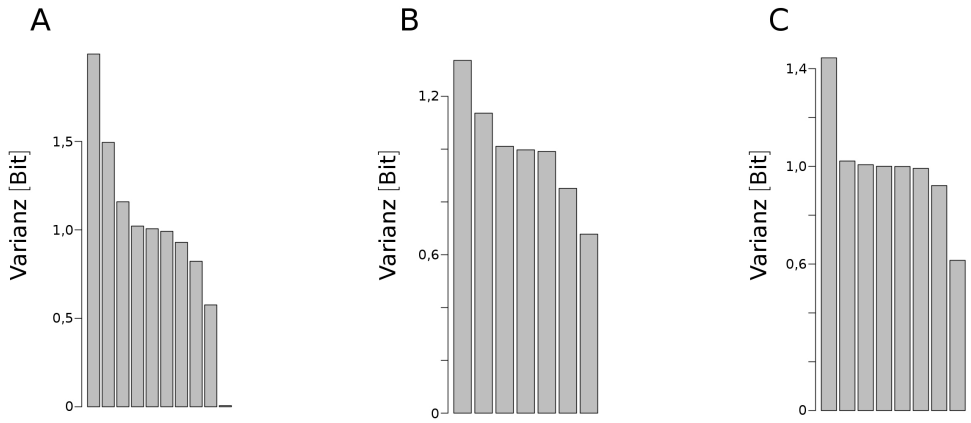

Abbildung 19: Aufgetragen sind die Varianzen der Hauptkomponenten aus den PCA. In (A) ist die PCA für alle Eigenschaften durchgeführt worden. (B) zeigt die PCA für möglichst überlappende Gruppen und in (C) wurden möglichst disjunkte Gruppen gewählt. Es wurden jeweils die vektorisierten 1Bit-MI-Matrizen verwendet, um die Matrix für die PCA zu bilden.

Die beobachtete Varianz der Hauptkomponenten der PCA kann als Maß des Informationsgehalts aufgefasst werden, wodurch in dieser Untersuchung deutlich wird, dass ein größerer Signalverlust in den überlappenden Gruppen (Abbildung 19 B) entsteht. Zwar ist der Informationsgehalt in der zweiten Hauptkomponente in den disjunkten Gruppen geringer als in den überlappenden, aber insgesamt wird mehr Information im disjunkten Gruppensatz erhalten. Aus diesem Grund wurden die anschließenden Untersuchungen mit den folgenden acht disjunkten Gruppen durchgeführt. `hydrophob`, `polar`, `aromatisch`, `winzig`, `positiv`, `aliphatisch`, `negativ`, `prolin`.

Die Koevolution für jede biochemische Eigenschaft aus den binär kodierten Alignments wurde positionsweise ausgewertet. Das heißt, es wurde ein binärer Vektor als Beschreibung der signifikanten Koevolution an jeder Position des Alignments für alle biochemischen Eigenschaften erstellt. Die Identifikation von signifikanten Werten wurde mit Hilfe der Z-Scores (Kapitel 3.2.2.5) durchgeführt. Ein signifikantes Signal besteht dann an einer Position P_i für eine biochemische Eigenschaft E_i, wenn an P_i ein Z-Score ≥ 4 [53] mit einer beliebig anderen Position P_j zu finden ist. In dem so erhaltenen Vektor wurden nicht korrespondierende Si-

gnale[7] gefunden. Das heißt, dass sich an P_i Koevolution für E_i und E_j finden, wobei die in den Eigenschaften E_i enthaltenen AS disjunkt von den AS der biochemischen Eigenschaft E_j vorliegen. Zum Beispiel wurde an der Position 10 des Lee-DS Koevolution der Eigenschaften `aliphatisch` und `aromatisch` identifiziert, was auf eine dramatische Änderung der Biochemie hinweist. Da die MI in den binär kodierten Sequenzen nicht differenziert, ob die Eigenschaft an der entsprechenden Position häufig oder gerade nicht häufig ist, wurden die Positionen mit nicht korrespondierenden Signalkombinationen näher untersucht. Dabei wurde die Häufigkeit der sich „widersprechenden" biochemischen Eigenschaften gegeneinander aufgetragen und so die Konserviertheit beider Eigenschaften an der jeweiligen Position überprüft. So ist der Widerspruch an Position 10 zum Beispiel falsch positiv, da das Signal aus dem häufigen Vorkommen der Eigenschaft `aliphatisch` und der gerade nicht vorkommenden Eigenschaft `aromatisch` entsteht. Bei der detaillierten Analyse aller nicht korrespondierenden Signale zeigt sich, dass der Großteil falsch positiv ist. So wurden in den verschiedenen Datensätzen immer nur ein geringer Prozentsatz als tatsächlich nicht korrespondierend identifiziert: Im Lee-DS sind nur 13 von 436 (ca. 3 %) tatsächlich nicht korrespondierend, im HIV-DS nur 17 von 1153 (ca. 1,5 %). Eine solch dramatische Änderung der biochemischen Eigenschaften könnte mit dem vom Proteaseinhibitor induzierten Selektionsdruck zusammenhängen. Um diesen Zusammenhang zu klären, wurde diese Analyse auch für den Teil der behandelten (HIV-DS (b)) und den der unbehandelten Sequenzen (HIV-DS (u)) des HIV-DS durchgeführt: Im HIV-DS (b) wurden 20 von 679 (ca. 3 %) als nicht korrespondierend bestätigt, im HIV-DS (u) nur 15 von 990 (ca. 1,5 %). Das Ergebnis zeigt, dass die Identifikation solcher Widersprüche nicht nur durch den Proteaseinhibitor induzierten Selektionsdruck entsteht, aber durch diesen mehr solcher Änderungen in der biochemischen Eigenschaft hervorgerufen werden können. In Tabelle 7 ist eine Aufstellung der identifizierten Residuen in den unterschiedlichen Datensätzen gezeigt.

Tabelle 7: Tabelle der Residuen, die jeweils signifikante Signale für disjunkte Gruppen biochemischer Eigenschaften aufweisen und in denen diese Eigenschaften auch zu mindestens 5 % konserviert vorliegen. Jeder Datensatz wurde getrennt untersucht. Durch $*$ markiert sind Positionen, die im entsprechenden Datensatz einen Widerspruch zeigen.

	10	16	20	33	37	63	71	72	82
Lee-DS					*	*	*	*	*
HIV-DS		*	*		*	*	*		*
HIV-DS (b)	*		*	*	*	*	*	*	*
HIV-DS (u)		*	*		*	*			

Die Änderung der biochemischen Eigenschaft kann für die Residuen 10, 16, 20, 33, 63, 71 und 82 auf einen durch die Behandlung mit PI induzierten Selektionsdruck zurückgeführt werden, da diese Residuen von Johnson et al. [74] als Residuen identifiziert wurden, die bei Behandlung

[7] Als nicht korrespondierend werden hier Signale an der selben Position für unterschiedlichen biochemischen Eigenschaften bezeichnet, die durch eine fehlende Schnittmenge an AS zunächst widersprüchlich erschienen. Als Beispiel ist hier eine hohes Signal der Eigenschaften `winzig` und `aliphatisch` an der selben Alignment-Position genannt. Solche Signale können durch eine tatsächliche Änderung in der Biochemie der AS entstehen, oder aber durch das vorhanden sein der einen z.B. `aliphatisch` und gerade der Abwesenheit der anderen Eigenschaft z.B. `winzig`.

mit PI am häufigsten mutieren. Für die Residuen 20, 37 und 63 kann der durch phylogenetische Effekte anliegende Selektionsdruck für eine solche Änderung der biochemischen Eigenschaften verantwortlich gemacht werden, da sich diese Residuen auch in dem von Liu et al. [81] beschriebenen phylogenetischen Varianzcluster (PhVC) befinden. Einzig das Residuum 72 ist in keiner der beiden Gruppen enthalten. Da sich das Signal an dieser Position jedoch im HIV-DS der Sequenzen aus behandelten Patienten finden lässt, kann vermutet werden, dass es sich hier um eine neu identifizierte Position handelt, die mit der Bildung von Resistenz assoziiert werden kann.

4.3.2 8Bit-cheMI

Die Interpretation der Ergebnisse aus den 1Bit-kodierten Sequenzen ist nicht trivial, da das 1Bit-Signal zunächst „widersprüchliche“ Signale zeigt, die nur mit näherer Analyse der Sequenz zu deuten sind. Nimmt man die Tatsache hinzu, dass eine AS über mehr als eine biochemische Eigenschaft verfügt, wird die Interpretation der 1Bit-cheMI noch komplizierter. Aus diesem Grund wurden die biochemischen Eigenschaften kombiniert und die Sequenzen der HIV-1 Protease unter Verwendung der disjunkten biochemischen Gruppen 8Bit-kodiert (siehe Kapitel 4.2.1).

Durch die kombinierte Kodierung der biochemischen Eigenschaften, die die AS-Sequenzen durch 8Bit-Vektoren ersetzt, sollten biochemisch koevolvierende Residuen identifiziert werden. Dabei war zu erwarten, dass sich biochemisch interagierende Residuen über besonders starke MI-Werte identifizieren lassen. Korreliert man die MI der 8Bit-Sequenzen mit der MI der AS-Sequenzen, ergibt sich ein Pearson-Koeffizient von 0,86. Dieser sehr viel höhere Wert stützt die Annahme, dass durch die 8Bit-Kodierung mehr Eigenschaften des Koevolutionssignals der AS-Sequenz abgebildet werden als in den Analysen der 1Bit-Kodierung. Dass die Korrelation jedoch nicht 1 ist, lässt den Schluss zu, dass trotz der Ähnlichkeit noch unterschiedliche Signale zu finden sind.

Die MI der AS-Sequenz und der 8Bit-Kodierung zeigte ein sehr starkes Signal im äußeren Dimerisierungsinterface der HIV-1 Protease (Residuen 1-13) [Daten nicht gezeigt], wobei die Entropie in diesem Bereich teilweise sehr gering ist. Um auszuschließen, dass das Gap-Symbol für die starke MI verantwortlich ist, wurde der Gehalt an nicht bekannten Aminosäuren (X) pro AS-Position bestimmt. Dabei wurde festgestellt, dass im Lee-DS bis zu 20 % der Sequenzen pro Position eine unbekannte AS tragen, wohingegen im HIV-DS nur 0,2 % der AS pro Position unbekannt sind (Abbildung 20 (A)). Um systematische Fehler bei der Sequenzierung ausschließen zu können, wurde das Auftreten der Platzhalter in den Sequenzen untersucht. Dazu wurde die Verteilung von einzelnen und konsekutiv aufeinander folgenden X in den Sequenzen betrachtet (Abbildung 20 B).

Aus der linearen Approximation der logarithmischen Auftragung in Abbildung 20 (B) lässt sich folgern, dass die X exponentiell verteilt sind und somit nicht durch systematische Sequenzierungsfehler entstanden sind. Bei Analyse des MI-Signals des Dimerisierungsinterfaces der HIV-1 Protease wurde festgestellt, dass die nicht sequenzierten AS zu hoher MI im Lee-DS führten. Diese Positionen sind ohne Betrachtung der Platzhalter fast vollständig konserviert und zeigen daher ein methodisch bedingtes Artefakt in der Berechnung der MI auf. Aus diesem Grund wurde in der weiteren Analyse die SUMI (Kapitel 3.2.1.1) berechnet.

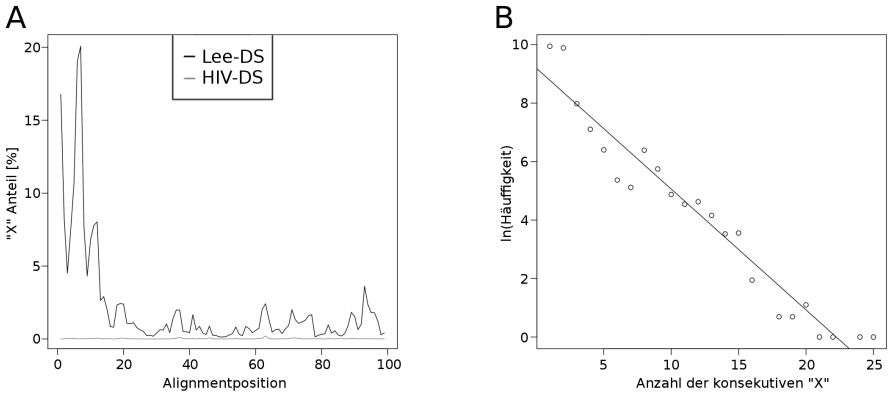

Abbildung 20: Prozentualer Anteil der X-Symbole an jeder Position des Alignments aus dem HIV-DS und den Sequenzen des Lee-DS (A) und die Verteilung der Häufigkeiten von konsekutiv auftretenden Sequenzierungsfehlern in den Sequenzen des Lee-DS, die durch ein X in der Sequenz markiert werden (B), logarithmische Häufigkeit aufgetragen über der Anzahl an konsekutiven X-Symbolen in der Sequenz aufgetragen. Die Steigung der angelegten Gerade in (B) beträgt -0,413.

Aufgrund der Dominanz der Entropie in der SUMI-Analyse (vergleiche Abbildung 21 (A) und (B)) wurden die Entropie und die SUMI getrennt voneinander untersucht. Um die Entropie in der SUMI zu vernachlässigen, wurde die Diagonale der SUMI-Matrix vor der SVD-Analyse und der Bestimmung der stärksten zehn Interaktionen manuell auf 0 gesetzt. Um einen generellen Eindruck der wichtigsten Interaktionspartner in der Struktur der HIV-1 Protease zu erhalten, wurden die aus der SVD-Analyse erhaltenen Eigenvektoren so rekombiniert, dass sie mindestens 80 % der SUMI-Matrix repräsentierten. Um die Signifikanz der SUMI zu bestimmen, wurden die Z-Scores berechnet. Da in den Sequenzen der HIV-1 Protease sehr hoch konservierte Positionen enthalten sind und dadurch die Varianz in der Berechnung des Nullmodells sehr klein wird, kommt es zu sehr hohen Z-Scores ($\gg$1.000). Abbildung 21 zeigt die SUMI-Matrix der Sequenzen des Lee-DS (mit und ohne Diagonale) und die Visualisierung dieser Daten anhand der Struktur der HIV-1 Protease.

Deutlich zu erkennen ist die Dominanz der Entropie auf der Diagonalen der MI-Matrix (vergl. Abbildung 21 (A) und (B)). Um die Ergebnisse der Koevolution zu visualisieren, wurde also die Diagonale vernachlässigt und für die weiteren Analysen manuell auf 0 gesetzt. Die Repräsentation der SUMI-Matrix durch den rekombinierten Eigenvektor auf der Struktur verdeutlicht den jeweiligen Einfluss des einzelnen Residuums in der SUMI-Matrix. Je dunkler also die Repräsentation des Residuums in Abbildung 21 (C), desto höher ist der Eintrag im rekonstituierten Eigenvektor (Kapitel 4.2.3), also mehr Einfluss in der SUMI-Matrix. Daraus erkennt man, dass starke Signale in den folgenden Bereichen des Moleküls zu finden sind. Starke koevolutionäre Prozesse finden sich also in der sogenannten *flap*-Region (Residuen 43-58 [140]), die aus β-Faltblättern besteht und das Molekül nach oben hin abschließt (Abbildung 21 (C)), sowie in der α-Helix (Residuen 86-94 [140]) und in den β-Faltblatt Regionen, die nicht zu der *flap*-Region gehören. Der Eigenvektor zeigt hohe Einträge hauptsächlich in Bereichen, die sich in Sekundärstrukturen befinden.

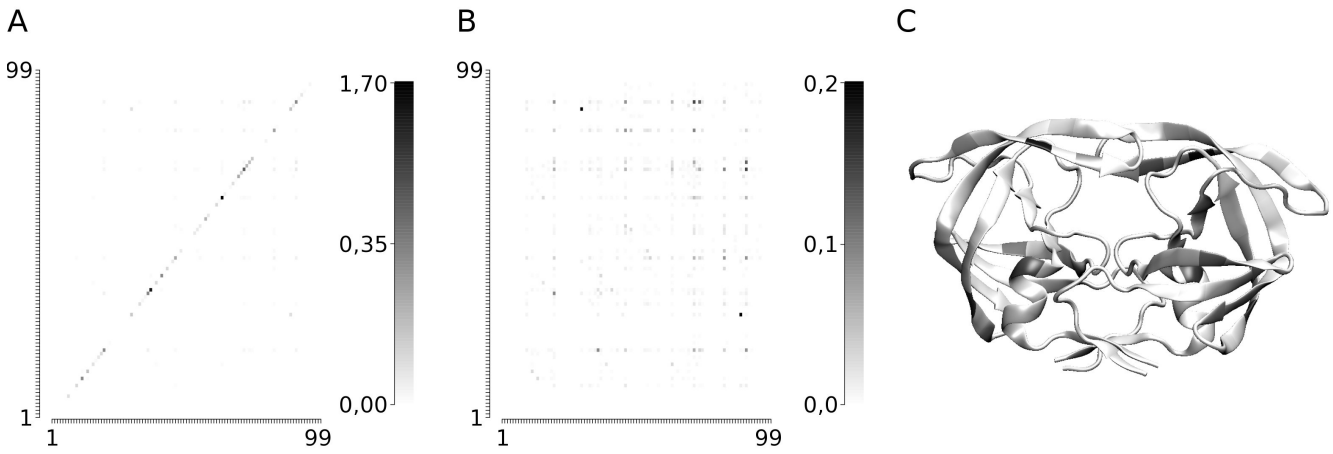

Abbildung 21: (A) Matrix der SUMI-Signale der 8Bit-Sequenzen aus dem Lee-DS. (B) Matrix aus (A), in der die Einträge der Diagonale manuell auf 0 gesetzt wurden. (C) Der kombinierte Eigenvektor (Kapitel 4.2.3), der 80 % der Matrix aus (B) repräsentiert, visualisiert auf der Kristallstruktur der HIV-1 Protease (PDB: 1KZK) mit Hilfe von VMD [71]. Der Verlauf des Farbcodes in (C) zeigt hohe Einträge in schwarz und niedrige in weiß. Die Einheit der MI-Matrizen und des kombinierten Eigenvektors sind Bit.

Ein vergleichendes Bild zwischen den zehn stärksten Signalen (Top10) der Sequenz-SUMI und den Top10 der 8Bit-SUMI ist in Abbildung 22 gezeigt.

Wie der Pearson-Koeffizient zwischen SUMI der AS-Sequenzen und den 8Bit-Sequenzen vermuten lässt, sind sich die Signale sehr ähnlich. Der Vergleich in Abbildung 22 zeigt jedoch schon in den Top10 Unterschiede in der Lage der koevolutionären Signale. So sind die Top10 der 8Bit-Sequenzen in dem von Liu et al. [81] beschriebenen phylogenetischen Cluster lokalisiert, wohingegen die Top10 der AS-Sequenz auch in anderen Bereichen zu finden sind. Auffällig sind hier vier Residuen, die durch die gleichen MI-Signale in der Sequenz und der 8Bit-Kodierung auffallen. Diese vier Residuen bilden ein Netzwerk zwischen dem *loop* in der *hinge*-Region, der Helix und dem β-Faltblatt in der Nähe der Helix. Im Lee-DS zeigt sich kein solch großes konserviertes Motiv im Vergleich der SUMI [Daten nicht gezeigt]. Es gibt drei koevolutionäre Interaktionen, die sich als gemeinsames Muster zwischen AS-Sequenz und 8Bit-Kodierung finden lassen (siehe Tabelle 8). Diese gemeinsamen Signale befinden sich zwischen der Helix und den Faltblättern der *flaps*.

Im allgemeinen Vergleich der Top10 SUMI der beiden Datensätze (HIV-DS und Lee-DS) zeigt sich ein deutlicher Unterschied, während sich die Top10 des HIV-DS im PhVC der Protease befinden, sind die Top10 des Lee-DS eher zur Mitte des Moleküls verschoben und liegen im von Liu et al. [81] gezeigten *drug resistance cluster* (DRC) [Daten nicht gezeigt]. Auffällig ist, dass in beiden Datensätzen die biochemische Koevolution hauptsächlich über lange Distanzen zu finden ist. Da biochemische Interaktionen oft zwischen benachbarten Residuen stattfinden, wurde die Distanzverteilung der schweren Atome zwischen den koevolvierenden Residuen berechnet. Dabei wurde für die interagierenden Residuen jeweils der kleinste Abstand der schweren Atome als Abstand der Residuen festgelegt. Das Minimum, das Maximum und der Mittelwert der

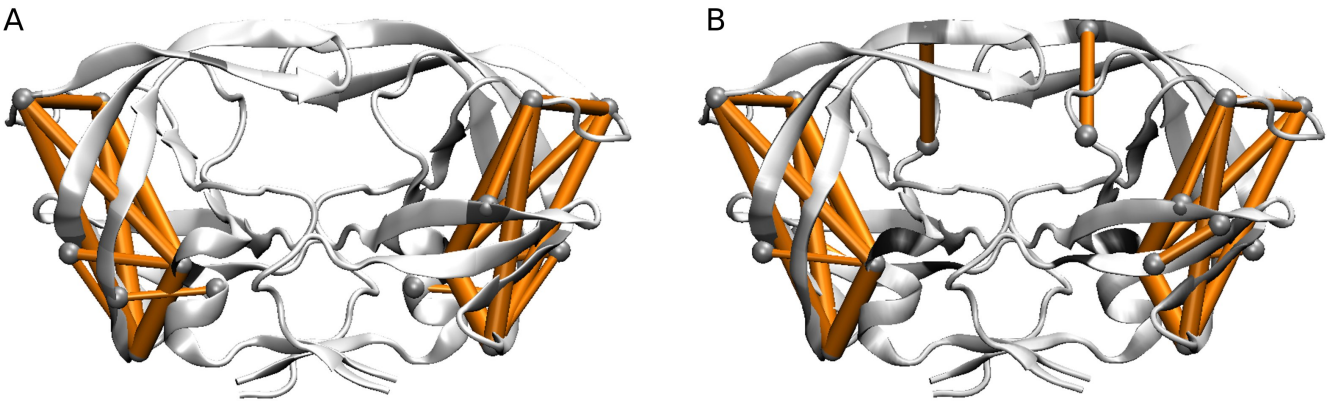

Abbildung 22: Die stärksten zehn SUMI-Signale aus dem HIV-DS. Die Kugeln repräsentieren die C_α-Atome der interagierenden (Verbindung zwischen den Kugeln) Residuen. In (A) die SUMI der 8Bit-Kodierung und in (B) die der AS-Sequenz. In Graustufen kodiert auf der Struktur, ist der kombinierte Eigenvektor aus der SVD-Analyse. Die stärke des SUMI-Signals entspricht der Verbindungsdicke.

Distanzen der Top10 Signale sind in Tabelle 9 gezeigt. Der Überblick der Distanzen zeigt, dass sich die stärksten koevolutionären Signale innerhalb der Protease zum Großteil nicht durch direkte Interaktion erklären lassen, da die Entfernung der Schweratome im Mittel deutlich von Distanzen abweicht, die auf eine direkte Interaktion schließen ließen, wie zum Beispiel 2 Å für eine Wasserstoffbrückenbindung. Zusätzlich ist im Vergleich der Distanzen von Signalen aus der AS-Sequenz mit den Signalen der 8Bit-Kodierung festzustellen, dass sich die Reichweite der koevolutionären Signale in ihren Kenngrößen kaum bis gar nicht unterscheidet. Die maximale Distanz zeigt, dass sich biochemisch koevolvierende Residuen finden, die auf entgegengesetzten Seiten des Monomers liegen (Durchmesser des Dimers ca. 53 Å). Die generelle Analyse der Signalstärke in Abhängigkeit von der Distanz zeigt eine leichte, negative Korrelation von $-0{,}12$ im Lee-DS und $-0{,}19$ im HIV-DS, was darauf hinweist, dass koevolutionäre Signale biochemischer Eigenschaften der HIV-1 Protease mit steigender Distanz zunehmen, statt abnehmen.

Diese Zusammenhänge lassen den Schluss zu, dass es sich bei den Top10 der 8Bit-SUMI hauptsächlich um indirekte biochemische Koevolution handelt. Diese koevolutionären Interaktionen können in der Packung der Viren-RNA im Capsid begründet liegen. Außerdem könnten die direkten Interaktionen so hoch konserviert sein, dass die biochemische MI hier 0 ist und somit nicht detektiert wird. Indirekte Interaktionen können auch durch eine Kette schwächerer Interaktionen entstehen, die die beiden stark koevolvierenden Residuen verbindet [21].

Die geringen Unterschiede der SUMI zwischen AS-Sequenz und 8Bit-Kodierung lassen vermuten, dass die von Liu et al. [81] gezeigten Cluster sich auch in der SUMI der 8Bit-kodierten Sequenzen identifizieren lassen. Dafür haben wir die SUMI des HIV-DS (b) und des HIV-DS (u) gegeneinander aufgetragen (siehe Abbildung 23 (A)), wobei zwei unterschiedliche Bereiche abzulesen sind. Zum Einen erkennt man Signale, die nur im HIV-DS (b) auftreten (im HIV-DS (u) eine SUMI von $\sim$0 haben), und zum Anderen solche, die im HIV-DS (u) ausgeprägter sind. Die Top10 beider Varianten sind in Abbildung 23 (B) auf der Struktur der HIV-1 Protease dargestellt.

Tabelle 8: Die zehn stärksten Interaktionen der beiden Datensätze aus den unterschiedlich kodierten Alignments, absteigend nach Interaktionsstärke sortiert.

Interaktionen			
Top10 Sequenz		Top10 8bit	
Lee-DS	HIV-DS	Lee-DS	HIV-DS
54 - 82	69 - 89	30 - 88	69 - 89
30 - 88	36 - 89	71 - 90	36 - 69
10 - 54	36 - 69	73 - 90	36 - 89
71 - 90	41 - 69	71 - 82	41 - 69
10 - 82	41 - 89	20 - 36	41 - 89
10 - 90	54 - 82	46 - 90	36 - 41
73 - 90	12 - 19	46 - 82	20 - 36
54 - 71	36 - 41	20 - 90	63 - 89
10 - 71	20 - 36	71 - 73	63 - 69
10 - 46	63 - 89	20 - 71	71 - 90

Tabelle 9: Kenngrößen der Distanzen (in [Å]) zwischen den Schweratomen zweier koevolvierender Residuen. Jeweils für die Top10 jedes Datensatzes. Es wurden nur die Entfernungen innerhalb der Ketten berechnet.

	HIV	HIV 8Bit	Lee	Lee 8Bit
Minimum	3,7	3,7	3,7	3,7
Mittelwert	10,2	10,4	11,9	10,3
Maximum	22,2	22,2	22,2	18,2

Um auszuschließen, dass es sich bei den beobachteten Signalen um Effekte handelt, die durch eine Überrepräsentation einer Spezies der HI-Viren (viele Isolate aus nur wenigen Patienten) oder durch statistisches Rauschen zustande kommen, wurden unterschiedliche Normierungsmethoden (Kapitel 3.2.2) verwendet und die normierte SUMI verglichen. Für alle verwendeten Normierungen ergeben sich ähnliche Muster der Koevolution, die einen Signalverlust im HIV-DS (b) und einen Signalgewinn im HIV-DS (u) aufweisen (siehe Abbildung 24).

Dabei ist besonders auffällig, dass es eine Teilmenge von vier Residuen gibt (37, 41, 69 und 89), die in jeder Normierung die gleichen Interaktionen aufweist. Diese Teilmenge ist auch im PhVC enthalten. Diese Beobachtung lässt darauf schließen, dass es sich bei diesen vier Residuen und den sie verbindenden koevolutionären Signalen um Schlüsselpositionen in der phylogenetischen Varianz der HIV-1 Protease handeln könnte.

4.3.2.1 Entropieanalyse

Mit Hilfe der Entropie haben wir die Variabilität des AS-Alphabets an jeder Position mit der biochemischen Variabilität an dieser Position verglichen. Besonders interessant sind dabei Positionen, die eine vergleichsweise große Entropie in der AS-Sequenz, aber eine geringe biochemische

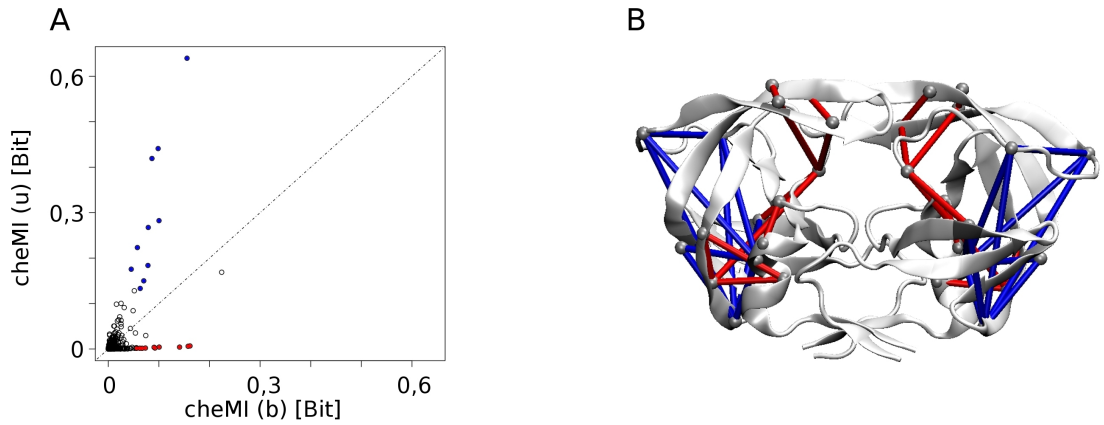

Abbildung 23: (A) zeigt den Scatterplot der Top10 der SUMI von HIV-DS (b) und HIV-DS (u). (B) Zeigt die in (A) markierten Interaktionen in der Struktur der HIV-1 Protease. In rot gekennzeichnet sind Interaktionen mit Signalgewinn in den Sequenzen der behandelten Patienten und in blau Interaktionen mit Signalverlust im Bezug auf das Signal im unbehandelten Datensatz.

Entropie aufweisen. Um die Entropien der AS und der 8Bit-Kodierung zu vergleichen, müssen sie normiert werden, da aufgrund der unterschiedlichen Alphabetgrößen ein direkter Vergleich nicht möglich ist. Die Bildung des Differenzsignals ist beschrieben durch:

$$\Delta H_i = \frac{\text{AS-}H_i}{\log_2 20} - \frac{\text{8Bit-}H_i}{\log_2 14} \tag{20}$$

Abbildung 26 zeigt das so berechnete Differenzsignal für den Lee-DS und den HIV-DS. An jeder Position, in der die AS-Entropie mehr Information beinhaltet als die der 8Bit-kodierten Sequenzen ergibt sich ein positives Signal. Die Identifikation interessanter Residuen erfolgte nach folgenden Kriterien: Das Differenzsignal ist größer als die biochemische Entropie und die Sequenzentropie ist nicht kleiner als 0,5. Als relevant wurden in den beiden Datensätzen die Residuen 13, 15, 35, 62, 77 und 93 identifiziert. Die Analyse wurde auch für HIV-DS (b) und HIV-DS (u) durchgeführt. Im Lee-DS und dem HIV-DS (b) wurde zusätzlich das Residuum 64 gefunden, das in keinem der anderen beiden Datensätze identifiziert wurde. Im nach behandelten und unbehandelten Sequenzen getrennten HIV-DS wurden die gleichen sechs Residuen gefunden. Zusätzlich wurde die Position 84 im HIV-DS (b) gefunden, die jedoch in keinem anderen Datensatz die geforderten Kriterien aufweist. Die in allen Datensätzen über das Differenzsignal identifizierten Residuen sind in Tabelle 10 gezeigt und mit der an dieser Position am stärksten konservierten biochemischen Gruppe (vergleiche Tabelle 5) in Tabelle 10 zusammengefasst. Die Residuen 13, 15, 63, 77 und 93 sind in Sekundärstrukturen der Protease zu finden und weisen alle eine Konservierung der biochemischen Gruppe XIV (`aliphatisch` und `hydrophob` vergleiche Tabelle 5) auf, nur das Residuum 35 befindet sich in einer *loop*-Struktur und ist konserviert für die Gruppe V (`negativ` und `polar`). Eine graphische Darstellung der Lage der sechs Residuen in der Protease ist in Abbildung 25 gezeigt.

Pan et al. [105] haben die Häufigkeit an auftretenden Punktmutationen in der Nukleotidsequenz der HIV-1 Protease untersucht und dabei Positionen identifiziert, an denen es häufig zu Punkt-

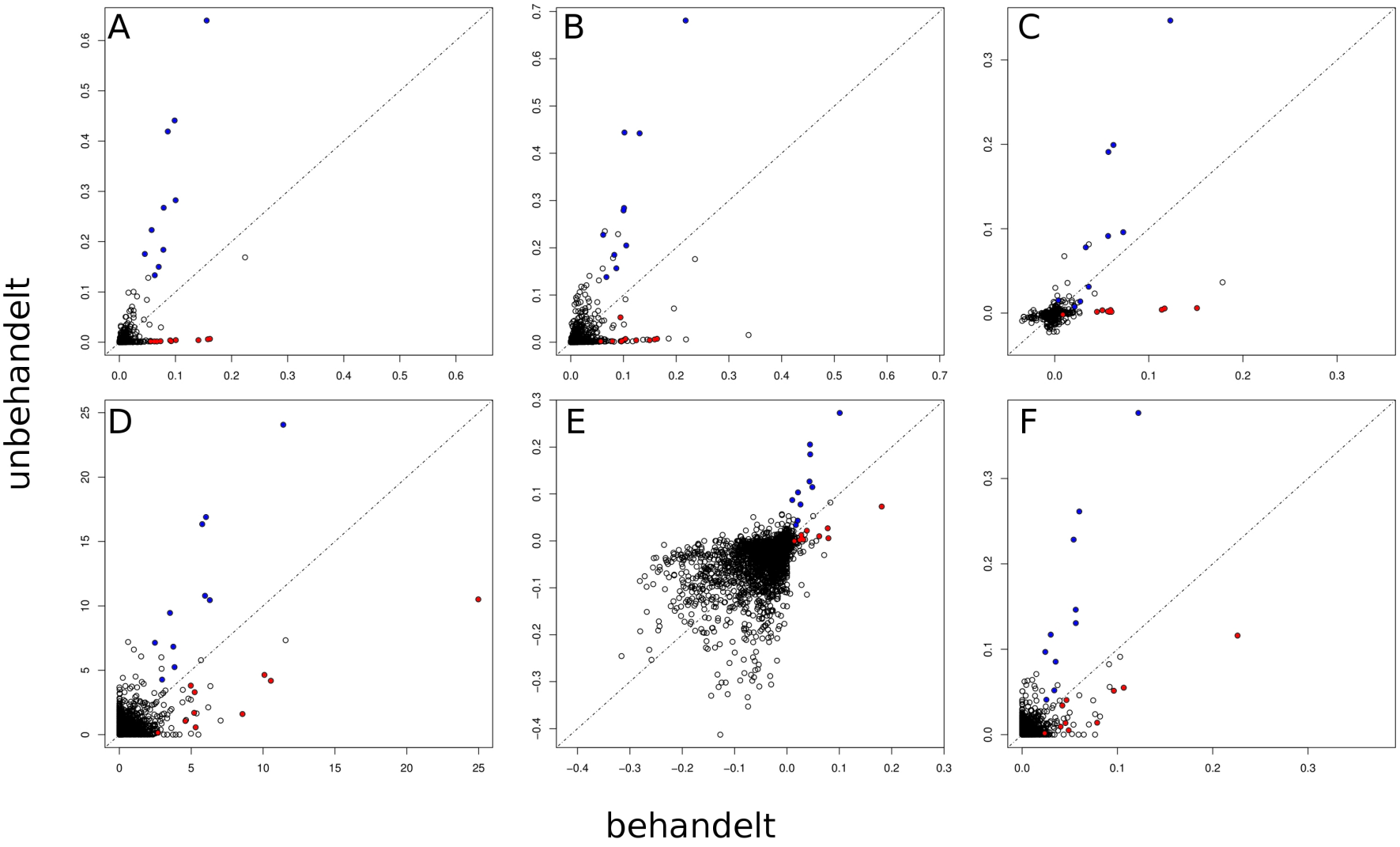

Abbildung 24: Auftrag der SUMI der HIV-DS (u) über HIV-DS (b). Farbig markiert sind die Top10 Interaktionen identifiziert aus (A). In (A) die Signale der 8Bit-Kodierung, (B) AS-Sequenz, (C) APC, (D) RCW, (E) gemeinsame Entropie (F) Spaltenentropie. (A), (B), (C), (E) und (F) jeweils in Bit angegeben.

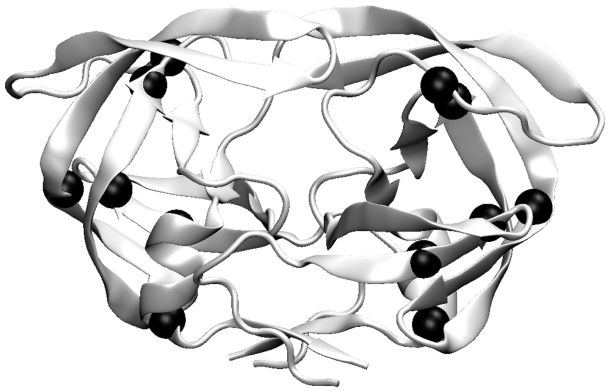

Abbildung 25: Residuen mit hoher Sequenzentropie und gleichzeitig niedriger biochemischer Entropie. Die Residuen wurden mittels des Differenzsignals identifiziert und sind hier durch schwarze Kugeln auf der C_α-Position repräsentiert.

mutationen kommt, die zu AS-Änderungen in der Proteinsequenz führen. Dafür haben Pan et al. ein Maß eingeführt, dass die Häufigkeit von nicht-synonymen Mutationen K_a mit der Häufigkeit von synonymen Mutationen K_s über den Quotienten K_a/K_s vergleicht. Wird der Quotient größer als 1, ergibt sich daraus das bevorzugte Einfügen von nicht-synonymen Mutationen. Vergleicht man die Ergebnisse von Pan et al. [105] mit den über das Differenzsignal identifizierten Residuen fällt auf, dass alle einen K_a/K_s größer als 1 zeigen. Das bedeutet zwar, dass an diesen Positionen bevorzugt eine Missense-Mutation eingefügt wird, die biochemischen Eigenschaften an diesen Positionen jedoch konserviert bleiben.

Die Residuen 77 und 93 finden sich in den am häufigsten mutierten Residuen wieder, die bei der Resistenzbildung beobachtet wurden [74]. Beide Residuen ändern jedoch nicht ihre biochemische Eigenschaft durch diese Mutation. Daraus lässt sich schließen, dass an diesen Positionen die biochemischen Eigenschaften besonders wichtig sind.

Tabelle 10: Relative Häufigkeit der am häufigsten vertretenen biochemischen Gruppe in Prozent. Es sind nur die Positionen gezeigt, die in allen Datensätzen eine hohe Konservierung aufwiesen.

Position	Gruppe	Lee-DS [%]	HIV-DS [%]	HIV-DS (b) [%]	HIV-DS(u) [%]
13	XIV	97,29	99,09	99,52	98,99
15	XIV	97,97	99,81	99,73	99,83
35	V	97,89	97,91	98,14	97,82
62	XIV	97,79	99,56	99,58	99,55
77	XIV	98,27	99,86	99,80	99,88
93	XIV	96,10	99,61	99,35	99,73

Im direkten Vergleich des Differenzsignals zwischen den beiden Datensätzen Lee-DS und HIV-DS zeigt sich ein sehr ähnliches Muster, das im Vergleich der Datensätze HIV-DS (u) und HIV-DS (b) große Unterschiede zeigt (siehe Abbildung 27). Vor allem im Bereich der *flaps* kommt es zu einer

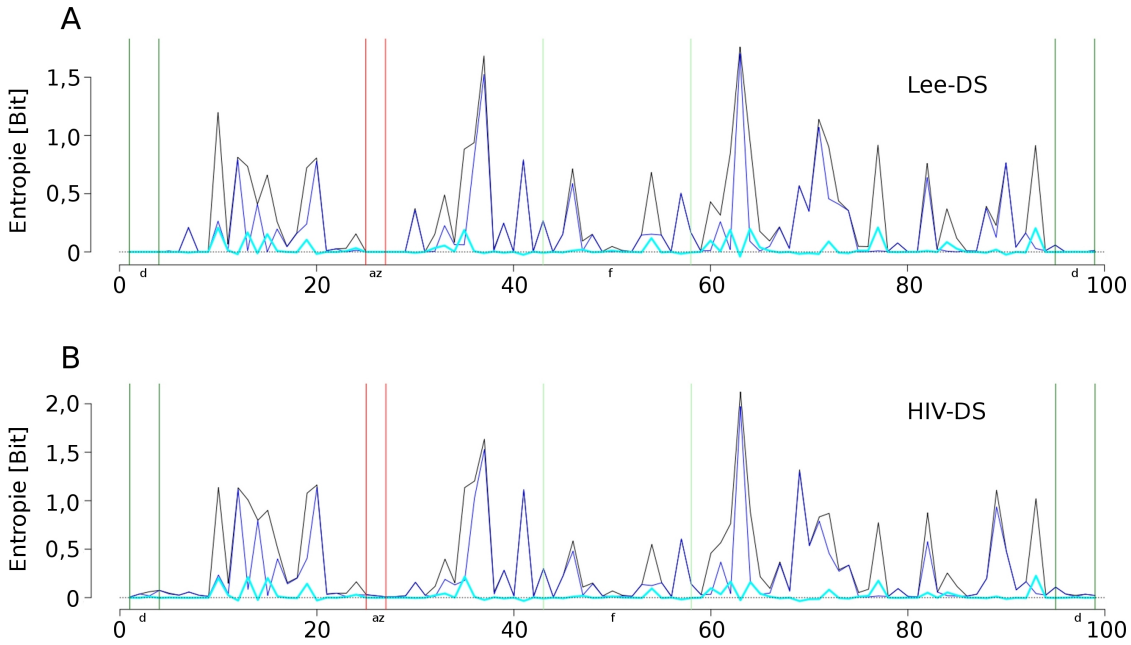

Abbildung 26: Vergleich zwischen der Sequenzentropie und der Entropie der 8Bit-Sequenzen jeweils für das Alignment aus dem Lee-DS (A) und HIV-DS (B). In schwarz ist die Sequenzentropie über der Alignmentposition gezeigt und in blau ist die biochemische Entropie aufgetragen. In cyan ist die Differenz der normalisierten Entropie-Vektoren gezeigt. Die vertikalen Linien kennzeichnen strukturelle Bereiche: d - Dimerisierungsinterface, f - *flaps* und az - aktives Zentrum.

hohen Entropieänderung in der Sequenz, die nur teilweise mit einer ähnlich starken Änderung in der biochemischen Entropie einhergeht. Das ist insofern interessant, als dass schon mehrfach gezeigt wurde, dass sich eine veränderte Flexibilität der *flap*-Region auf die Sensitivität des Moleküls auswirken kann und auch bei der Resistenzbildung beobachtet wird [79, 5, 62, 102].

Durch die Verbindung des Differenzsignals mit der maximalen Koevolution an jeder Position im Vergleich zwischen AS-Sequenzen und 8Bit-Kodierung lässt sich der Selektionsdruck auf biochemischer Ebene verdeutlichen. Der Zusammenhang zwischen den Signalen in den vier unterschiedlichen Datensätzen ist in Abbildung 28 gezeigt. Wie deutlich zu erkennen ist, nimmt die Koevolution der 8Bit-Kodierung mit steigendem Differenzsignal ab und korreliert immer weniger mit den maximalen MI-Werten aus den AS-Sequenzen.

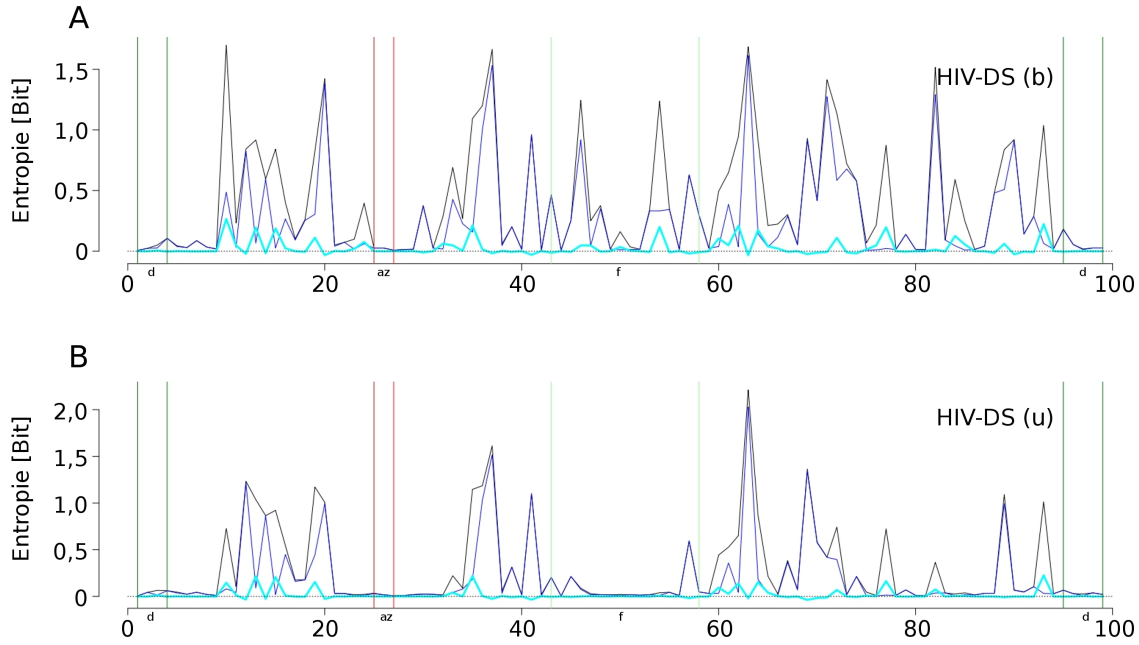

Abbildung 27: Vergleich zwischen der Sequenzentropie und der Entropie der 8Bit-Sequenzen jeweils für den HIV-DS (b) (A) und den HIV-DS (u) (B). In schwarz ist die Sequenzentropie über der Sequenzposition gezeigt und in blau ist die biochemische Entropie aufgetragen. Die Differenz der normalisierten Entropien ist in cyan gezeigt. Die vertikalen Linien kennzeichnen strukturelle Bereiche: d - Dimerisierungsinterface, f - *flaps* und az - aktives Zentrum.

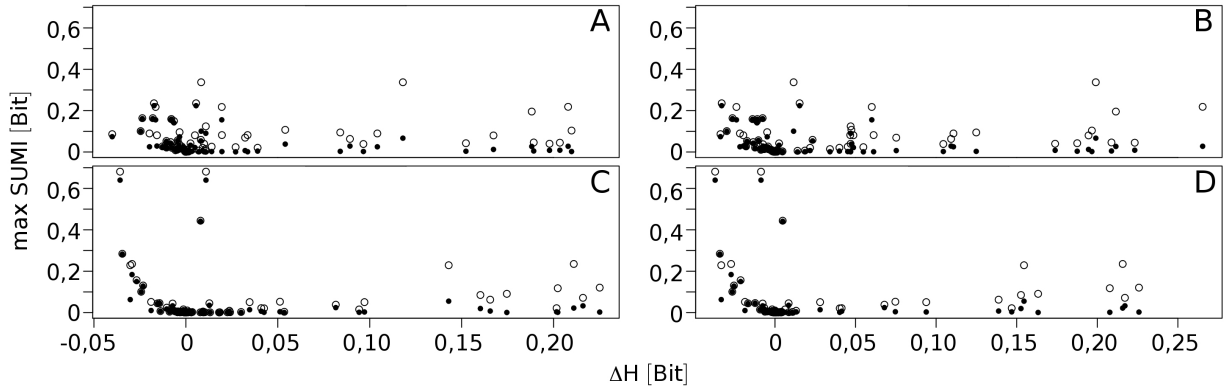

Abbildung 28: Aufgetragen sind die jeweils maximalen SUMI-Signale gegen die Differenzsignale der Entropie-Analyse jeder Position. In (A) ist die Analyse für den Lee-DS gezeigt, für den HIV-DS (b) in (B), für den HIV-DS in (C) und für den HIV-DS (u) in (D). Die Analyse wurde jeweils für die Signale der SUMI der AS-Sequenzen (offene Kreise) und für die 8Bit-Sequenzen (schwarze Punkte) durchgeführt.

4.4 Diskussion

4.4.1 cheMI

4.4.1.1 1Bit-MI

Mit der Analyse der 1Bit-kodierten Sequenzen ist es möglich, Residuen zu identifizieren, die bevorzugt unter Änderung der biochemischen Eigenschaft mutieren. Das deutet darauf hin, dass die biochemischen Eigenschaften an diesen Positionen keine gravierenden Änderungen der Funktionalität der Protease verursachen und das Virus weiter proliferieren kann. Dass diese Residuen für die Ausbildung einer Resistenz mitverantwortlich sein könnten, lässt sich auf Grund der von Johnson et al. [74] identifizierten Residuen vermuten. Die Residuen 37 und 72 gehören zwar nicht zu den am häufigsten selektierten Positionen, die zur Resistenzbildung in der HIV-1 Protease beitragen, könnten jedoch als Hinweise auf solche gewertet werden, was sich auf zwei Tatsachen stützen lässt. Zum Einen vergrößern sich Datenbestände, wie zum Beispiel die von Johnson et al. [74], kontinuierlich, sodass sich immer mehr Residuen mit der Resistenzbildung in Verbindung bringen lassen. Zum Zweiten lassen sich die beiden genannten Residuen nicht dem von Liu et al. [81] gezeigten PhVC zuordnen, wodurch die beobachtete Änderung der biochemischen Eigenschaft nicht auf den phylogenetischen Selektionsdruck zurückzuführen ist.

Die hohe Rate an falsch positiv identifizierten Signalen ist durch die fehlende „Richtung“ der Analyse begründet. Dabei ist „Richtung“ in diesem Zusammenhang als konserviert oder gerade nicht konserviert für eine bestimmte biochemische Eigenschaft zu verstehen. Diese Einschränkung sollte in weiteren Anwendungen berücksichtigt werden. Trotz dieser Einschränkung können mit der Analyse der 1Bit-kodierten Sequenzen, Positionen eines Moleküls identifiziert werden, die eine interessante biochemische Änderung im untersuchten Datensatz aufweisen.

4.4.1.2 8Bit-MI

Die MI-Analyse der 8Bit-kodierten Sequenzen zeigt zwei Cluster, die dem PhVC und dem DRC von Liu et al. [81] sehr ähnlich sind. Dabei ist die Übereinstimmung des PhVC wesentlich deutlicher, als die des DRC. Diese Übereinstimmung lässt vermuten, dass die koevolutionären Signale der AS-Sequenzen zumindest teilweise durch biochemische Interaktionen zu erklären sind. Dabei ist anzumerken, dass auch die Koevolution der biochemischen Eigenschaft am stärksten über lange Reichweite stattfindet und nicht wie zunächst erwartet in direktem Kontakt. Die „fehlenden“ koevolutionären Signale zwischen Residuen in Interaktionsreichweite könnten durch starke Konservierung der Eigenschaften zustande kommen. Wie bereits erwähnt, konnten wir das PhVC-Cluster von Liu et al. [81] auch in den biochemischen Eigenschaften finden. Die sehr unterschiedlichen Ergebnisse des DRC können durch die unterschiedlichen Datensätze erklärt werden, die in den Studien genutzt wurden. Im Gegensatz zu Liu et al. [81] konnten wir zeigen, dass in den Clustern die Signalstärke variiert. Die Abnahme der Signalstärke im PhVC gegenüber dem DRC im HIV-DS (b) lässt darauf schließen, dass die biochemischen Interaktionen der Resistenzmutationen im Vergleich mit denen des PhVC eine übergeordnete Rolle spielen.

Das Differenzsignal der hier verwendeten Datensätze zeigt eine gute Übereinstimmung im direkten Vergleich der Datensätze Lee-DS und HIV-DS. Im Vergleich der HIV-DS (b) und HIV-DS (u) zeigen sich deutliche Unterschiede hauptsächlich in der *flap*-Region, was auf zugrundeliegende biochemische Mechanismen schließen lässt, die in diesem Bereich für Resistenzmechanismen verantwortlich sind. Die Interaktionen dieses Bereichs weisen deutliche Einschränkungen der Variabilität auf biochemischer Ebene auf, im Gegensatz zu der relativ hohen AS-Variabilität.

Auch Veränderungen in der Variabilität in der 8Bit-Kodierung und der AS-Variabilität ergeben Hinweise auf interessante Residuen bei der Resistenzgewinnung des Moleküls.

Die Residuen, identifiziert durch das Differenzsignal, gehören in der HIV-1 Protease hauptsächlich der biochemischen Gruppe XIV an. Diese Gruppe ist im gesamten Molekül sehr verbreitet, da ca. 33 % aller Residuen zu mindestens 67 % für diese Gruppe konserviert sind. Diese Beobachtung lässt zwar keinen direkten Schluss über die so gefundenen Residuen zu, zeigt aber die große Relevanz der Hydrophobizität innerhalb des Moleküls. Das häufige Auftreten der Gruppe XIV zeigt weiterhin, dass auch aliphatische Wechselwirkungen wichtig zu sein scheinen. Die beiden nur in jeweils einem der Datesätze identifizierten Residuen 64 und 84 gehören mit den Residuen 93 und 77 zwar zu den von Johnson et al. [74] identifizierten Resistenzmutationen, ändern aber dabei ihre biochemische Gruppe nicht. Diese Tatsache könnte ein Hinweis darauf sein, dass hier andere Faktoren als die Biochemie eine Rolle in der Resistenzbildung spielen. Diese Vermutung wird auch durch das Residuum 82 deutlich, das als Hauptmutation von Johnson et al. [74] aufgezeigt wurde. Die Identifikation von Residuen über das Differenzsignal zeigt also Positionen mit hohem biochemischem Selektionsdruck, die in ihrer AS-Variabilität jedoch weniger beschränkt sind.

In Abbildung 28 lässt sich ein biochemischer Selektionsdruck durch das hohe Differenzsignal ableiten. Die gleichzeitig beobachtete Koevolution auf biochemischer Ebene ist bei hohem Differenzsignal immer geringer als die maximale koevolutionäre Interaktion auf AS-Ebene. Interessanterweise existieren aber auch negative Differenzsignale, die darauf hindeuten, dass hier die biochemische Koevolution wesentlich stärker ist als die der AS. Somit ergibt sich für die in Abbildung 28 gezeigten Signale die Unterscheidung in biochemisch evolvierte Signale, deren AS noch variieren kann, und Positionen, an denen weder biochemische Evolution noch die Evolution der AS-Sequenz abgeschlossen sind.

Alle hier gezeigten Untersuchungen der 8Bit-kodierten Datensätze zeigen sehr deutliche Unterschiede zwischen dem Lee-DS und den Ergebnissen des HIV-DS. Trennt man den HIV-DS jedoch in HIV-DS (u) und HIV-DS (b), werden zwei Dinge deutlich. Zum Einen fällt die Ähnlichkeit der Ergebnisse für dem Lee-DS und dem HIV-DS (b) auf. Diese Beobachtung weist darauf hin, dass der Lee-DS fast vollständig aus Sequenzen behandelter Patienten stammt. Diese Vermutung lässt sich nicht eindeutig beweisen, da der Lee-DS zwar aus ausschließlich behandelten Patienten stammt, deren Behandlungsstatus jedoch nicht identifizierbar ist.

Zum Anderen zeigen die ähnlichen Ergebnisse der MI-Signale zwischen HIV-DS und HIV-DS (u), dass durch den hohen Anteil (ca. $^2/_3$) an unbehandelten Sequenzen das Signal der behandelten Sequenzen stark maskiert wird. Ganz anders verhält es sich auf Ebene des Differenzsignals, das es ermöglicht, die Signalunterschiede zu den Sequenzen behandelter Patienten zu identifizieren. Diese Beobachtung zeigt, dass eine getrennte Analyse von Entropie und MI sinnvoll ist,

um Effekte, die durch Behandlung entstehen, auf biochemischer Ebene zu verdeutlichen. Die Interpretation nur eines Signals in 8Bit-kodierten Sequenzen gibt zunächst nur vage Hinweise, kann aber in Verbindung mit den Signalen aus den AS-Sequenzen und der Interpretation aller Analysen auf interessante Residuengruppen des Moleküls hinweisen.

5 Biophysikalische Netzwerke

5.1 Einleitung

Die strukturelle Aufklärung des Ribosoms hat zum seinem funktionellen Verständnis beigetragen [114, 12]. Die erste MD-Simulation des Ribosoms wurde 2006 von Cui et al. [34] durchgeführt. Untersuchungen von Antibiotika-Bindestellen können mit Hilfe von MD-Simulationen durchgeführt werden, wie Aleksandrov et al. am Beispiel von Tetracyclin gezeigt haben [2]. Des Weiteren konnte mit Hilfe von Analysen elastischer Netzwerkmodelle die Dynamik des ribosomalen Tunnels aufgeklärt werden, der den Weg für des naszierende Protein aus dem Ribosom darstellt [77]. Auch die Evolution des Ribosoms konnte mit Hilfe der Bioinformatik untersucht werden [117]. Durch die verfügbaren Kristallstrukturen, die das Ribosom mit gebundenem Aminoglykosid-Antibiotika zeigen konnte in Verbindung mit MD-Simulationen, die Funktionsweise von Aminoglykosid-Antibiotika detaillierter aufzuklärt werden [118].

Durch die Verwendung von reduzierten Modellen konnte von Hamacher et al. [65] an der kleinen ribosomalen Untereinheit die Assemblierung des Komplexes durch computergestützte Analyse nachvollzogen werden. Mit Hilfe der *self consistent pair contact probability approximation* [95] (SCPCP) (Kapitel 5.2.2) wurde die 30 S Untereinheit des Ribosoms untersucht. Dafür wurden die Berechnungen der Bindungsenergie sowohl für die gesamte Untereinheit als auch für die einzelnen Molekülketten durchgeführt. Anschließend wurden die Bindeenergien bestimmt, nachdem eine bzw. zwei Molekülketten entfernt wurden. Durch diese Methode können Differenzen der Bindeenergie berechnet werden, die Aufschluss darüber geben, welche Molekülkette die Bindung welcher anderen Kette beeinflusst [65]. Diese Analyse von Hamacher et al. hat gezeigt, dass mit Hilfe der SCPCP thermodynamische Berechnungen ermöglicht werden, die zur Bestimmung von Bindeenergien notwendig sind.

Im Folgenden soll diese Methode auf die große ribosomale Untereinheit transferiert werden.

5.2 Methoden

5.2.1 Biophysikalische Netzwerke

Um die biophysikalischen Eigenschaften eines Proteins aufzuklären, haben Brooks et al. und Go et al. die Normalmodenanalyse (NMA) etabliert [18, 54]. Die NMA gibt Aufschluss über die Fluktuationen der atomaren Lokalisation und kann so funktionelle Moden der Proteindynamik aufzeigen. Dabei werden die Raumkoordinaten jedes Atoms in einer Proteinstruktur als nahe ihres Gleichgewichtszustandes betrachtet, das heißt die Raumkoordinate der Kristallstruktur wird als ein Mittel der tatsächlichen fluktuierenden Position gewertet. Durch die Berechnung des Potentials der Strukturen und der aus dem Potential abgeleiteten, korrelierten Bewegungen lassen sich so elementare Aussagen über die Molekülbeweglichkeit treffen. Eine vereinfachte Form der NMA ist die Betrachtung von elastischen Netzwerkmodellen (ENM), die auf einer Residuen basierten Betrachtung des Netzwerks beruhen [59, 9]. Diese Vereinfachung wurde inspiriert durch die ganz atomistischen Betrachtungen von Tirion [131] unter Verwendung von uniformen harmonischen Potentialen. Die Darstellung der Proteinstruktur als Netzwerk mit harmonischen Interaktionen ist als Schema in Abbildung 29 gezeigt.

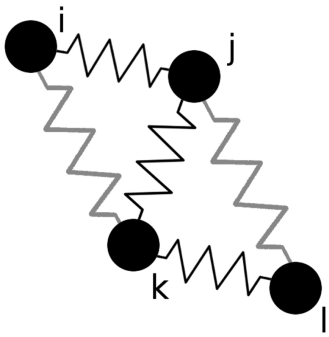

Abbildung 29: Schematische Darstellung eines elastischen Netzwerkmodells. Die einzelnen Atome (schwarze Kreise) sind durch unterschiedliche Interaktionen miteinander verbunden. Dabei sind hier kovalente Bindungen (schwarze Federn) und nichtkovalente Interaktionen (graue Federn) abgebildet.

Zu den ENM gehören die anisotropen Netzwerkmodelle (ANM) [8] und die Gauß'schen Netzwerkmodelle (GNM) [9]. Dabei stellen die GNM die einfachste Form der ENM dar. Die GNM-Analyse betrachtet ausschließlich topologische Randbedingungen der Struktur, bei der das Protein residuenbasierend als Netzwerk dargestellt wird. Dabei werden die C_α-Atome der Residuen als Knoten des Netzwerkes abgebildet, die durch uniforme harmonische Interaktionen miteinander verbunden sind. Die Interaktion zweier Residuen i und j beruht dabei allein auf dem Abstand ΔR_{ij} der C_α-Atome dieser Residuen und wird als vorhanden definiert, wenn ein bestimmter Abstand ΔR_c unterschritten wird. Der Abstand wird dabei über die Fluktuation R_i eines einzelnen Residuums i von der Ausgangsposition R_i^0 beschrieben, der durch Subtraktion den Distanzvektor $\Delta R_i = R_i - R_i^0$ ergibt. Der Abstand ΔR_{ij} für zwei Residuen ergibt sich dann aus $\Delta R_j - \Delta R_i$. Unter der Annahme, dass diese Fluktuationen isotrop und gaußverteilt vorliegen, ergibt sich das Potential des Netzwerks mit N Knoten aus den Raumkoordinaten der C_α-Atome nach folgendem Zusammenhang:

$$V_{\mathrm{GNM}} = \frac{\gamma}{2}\left[\sum_{i,j}^{N} \Gamma_{ij} \underbrace{(\Delta R_j - \Delta R_i)^2}_{=:\Delta R_{ij}}\right] \tag{21}$$

γ ist die Federkonstante, die die Stärke der Interaktion definiert. Γ_{ij} ist die Kirchhoffmatrix, die eine Art Abbildung der Kontakttopologie darstellt, wobei der Eintrag $\Gamma_{ij} = -1$ ist, wenn die Residuen i und j in Kontakt stehen, $\Delta R_{ij}^{\mathrm{PDB}}$[8] also kleiner oder gleich R_c ist. Die Diagonaleinträge

[8] $\Delta R_{ij}^{\mathrm{PDB}}$ sist hier der Abstand der Residuen i und j innerhalb der Kristallstruktur.

von Γ entsprechen der negativen Summe der jeweiligen Matrixzeile. Zusammenfassend ist die Kirchhoffmatrix wie folgt definiert:

$$\Gamma_{ij} = \begin{cases} -1, & \text{wenn } i \neq j \text{ und } R_{ij} \leq R_c \\ 0, & \text{wenn } i \neq j \text{ und } R_{ij} > R_c \\ -\sum_{j,j\neq i}^{N} \Gamma_{ij}, & \text{wenn } i = j \end{cases} \tag{22}$$

Aus diesen Überlegungen und der Annahme von isotropen Bewegungsfluktuationen kann die mechanische Korrelationsmatrix im thermodynamischen Ensemble wie folgt berechnet werden [64]:

$$C_{ij} := \frac{3k_B T}{\gamma} \widetilde{\Gamma}_{ij}^{-1} \tag{23}$$

$\widetilde{\Gamma}^{-1}$ ist die Moore-Penrose-Pseudoinverse [96, 107] von Γ. Die Invertierung muss hier nummerisch aufwendiger durchgeführt werden, da das GNM-Modell welches Γ erzeugt, einen Symmetriefreiheitsgrad aufweist und somit Γ einen sigulärwert gleich Null hat – also nicht exakt invertierbar ist. In dieser Arbeit wurde die GNM-Analyse an vier Strukturen der ribosomalen Untereinheiten für Proteine und RNA durchgeführt. Dazu wurden zwei dissoziierte 50 S und 30 S Ribosomen aus der PDB verwendet; dabei handelt es sich im Speziellen um die 50 S Untereinheit von *T. thermophilus* (PDB-Code 1VSA) und die 50 S Untereinheit von *E. coli* (PDB-Code 2AWB), sowie den 30 S Untereinheiten von *T. thermophilus* und *E. coli* (PDB-Codes 1J5E und 2AVY). Alle PDB-Strukturen wurden manuell auf Alternativeinträge für Residuen untersucht und falls vorhanden wurden alle alternativen Einträge entfernt. Die Sequenzen der einzelnen rRNAs und der Proteine wurden jeweils aus beiden PDB-Strukturen extrahiert und mit `clustalw` [130, 28, 78] aligniert. Diese Alignments wurden mit Hilfe der `Jalview`-Software [31, 134] manuell nachbereitet. Aus den so prozessierten Alignments wurden Übersetzungstabellen erstellt, die die korrespondierenden Kontakte der ribosomalen Untereinheiten der beiden Spezies enthalten.

Für die Berechnung der GNMs wurden aus den PDB-Strukturen die Koordinaten für die schweren Atome (C, N, O, P, S) jedes Residuums extrahiert und die Distanzen innerhalb jeder Untereinheit bestimmt. Als Grundlage für die Berechnung der GNMs wurde die Kontaktmatrix mit einem Wert von 6.75 Å für R_c erstellt [112]. Für die berechneten Modelle wurden anschließend die inversen Korrelationsmatrizen errechnet und als Wildtyp-Matrix C^w gespeichert. Anschließend wurde für jedes der Modelle jeder Kontakt einzeln durch die Modifikation der Kirchhoffmatrix „ausgeschaltet“ und die Berechnung erneut durchgeführt. Die Unterschiede in der Korrelationsmatrix der mutierten Matrix $C^{o_{ij}}$ zur Wildtyp-Matrix wurden mit Hilfe der Frobeniusnorm (Gleichung 24) bestimmt.

Die Frobeniusnorm (FN) ist definiert als eine euklidische Matrixnorm. Für die Berechnung der FN in den GNM-Analysen wurden alle Einträge k und l der Wildtyp-Matrix mit allen Einträgen der mutierten Matrix verglichen. Die FN in den GNM-Analysen stellt sich also wie folgt dar:

$$\mathrm{FN} = \sqrt{\sum_{k,l} (C_{kl}^{w} - C_{kl}^{o_{ij}})^2} \tag{24}$$

Die kovalenten Bindungen des Peptidrückgrats wurden nicht zu berücksichtigt.

5.2.2 *self consistent pair contact probability approximation* (SCPCP)

Die Methode der *self consistent pair contact probability approximation* (SCPCP) [95] wurde von Micheletti et al. 2001 vorgestellt. Dabei handelt es sich um eine Methode, die es ermöglicht, die Gleichgewichtseigenschaften von Proteinen approximativ zu ermitteln. Hier wird aus einer PDB-Struktur die freie Bindungsenergie berechnet. Die Berechnung basiert auf der Reduktion des Proteins auf seine C_α-Atome, beziehungsweise der P-Atome, wenn es sich bei dem biologischen Molekül um eine RNA oder DNA handelt. Die Reduktion der AS auf ihre C_α-Atome stellt die Knoten des so aufgebauten Netzwerks dar, das über das Rückgrat durch harmonische Interaktionen verbunden ist. Die Kontakte innerhalb des Moleküls werden über einen Distanzschwellenwert von $R_c = 3{,}75\,\text{Å}$ definiert und im Netzwerk durch weitere harmonische Interaktionen repräsentiert [65]. Über die Summe der harmonischen Interaktionen kann dann die potentielle Energie des Moleküls berechnet werden. Dabei ist die Summe der harmonischen Interaktionen in der so genannten Hamiltonischen Form ausgedrückt:

$$H = \frac{T}{2}\sum_{i=1}^{N-1} K_{i,i+1}\Xi_{i,i+1}\tilde{X}_{i,i+1}^2 - \sum_{i,j=1}^{N}\frac{\Delta_{ij}\kappa_{ij}}{2}\left[R^2 - \tilde{X}_{ij}^2\right]\cdot\Theta(R^2 - \tilde{X}_{ij}^2) \tag{25}$$

dabei ist

$$\tilde{X}_{ij}^2 = (\vec{r}_{ij} - \vec{r}_{ij}^0)^2 = (\Delta\vec{r}_{ij} - \Delta\vec{r}_{ij}^0)^2. \tag{26}$$

Der Kontakt Δ_{ij} innerhalb des Moleküls ist genau dann vorhanden, also 1, wenn der Distanzschwellenwert der schweren Atome der beobachteten Moleküle kleiner ist als R_c. Ob der beobachtete Kontakt eine kovalente Bindung über das Rückgrat des Moleküls darstellt, wird über den Term $\Xi_{i,i+1}$ dargestellt, der genau dann 1 ist, wenn es sich um einen kovalenten Kontakt handelt. Bei K handelt es sich um die Definition von Pseudobindungen, wohingegen durch κ die AS-spezifische Bindung moduliert wird. Die genaue Parametrisierung dieser spezifischen Bindungsstärken wurde von aus den Resultaten von Hamacher et al. [65] übernommen. Da die Integration der Hamiltonischen Form (Gleichung 25) zum Beispiel über die MD-Simulation des Molekülkomplexes sehr zeitaufwendig ist, haben Micheletti et al. [95] eine Form gefunden, die die Berechnung deutlich beschleunigt und die Simulation überflüssig macht. Diese Methode von Micheletti et al. beruht auf der Kontaktwahrscheinlichkeit

$$p_{ij} = \left\langle \Theta\left\langle R^2 - \tilde{X}_{ij}^2\right\rangle\right\rangle_H,$$

die den thermodynamischen Erwartungswert der kontaktdefinierenden Funktion darstellt. Dadurch kann der Term in Gleichung 25 durch p_{ij} ersetzt werden und durch diese *meanfield*-Approximation Gauß ähnliche Integrale in der Zustandssumme erhalten. Durch das schnelle

Konvergieren von p_{ij} ist es dann möglich, die Berechnung iterativ durchzuführen. Dabei folgt die Iteration für den Schritt $n+1$ dem Schema in Gleichung 27-29.

$$p_{ij}^{n+1} = \left(\frac{3}{2}; \frac{R^2}{2G_{ij}^{(n)}} \right) \tag{27}$$

$$G_{ij}^{(n)} = M_{ii}^{(n)} + M_{jj}^{(n)} - M_{ij}^{(n)} - M_{ji}^{(n)} \tag{28}$$

$$M_{ij}^{-1(n)} = \begin{cases} K_{ij}(\Xi_{i,i+1} + \Xi_{i,i-1}) + 2\sum_l \Delta_{il}\kappa_{il}p_{il}^{(n)}/T & i = j \\ -2p_{ij}^{(n)}\Delta_{ij}\kappa_{ij}/T - K_{ij}(\delta_{i,j+1} + \delta_{i,j-1}) & i \neq j \end{cases} \tag{29}$$

Daraus lässt sich nun die freie Energie durch den folgenden Zusammenhang ableiten:

$$F/T = -\frac{3}{2}(N\ln(2\pi) + \ln(detM)) - \frac{R^2}{2T}\sum_{ij}\Delta_{ij}\kappa_{ij}p_{ij} \tag{30}$$

5.3 Resultate

5.3.1 Kontaktannotation via GNM

Um die Kontakte des Ribosoms zu annotieren, die wichtig für die Moleküldynamik sind, wurden GNM-Studien an Strukturen der kleinen und großen Untereinheit des Ribosoms durchgeführt. Außerdem sollte untersucht werden, ob in verschiedenen Organismen bei einem makromolekularen Komplex mit so hoch konservierter Funktion eine Übereinstimmung hinsichtlich dieser mechanistischen Eigenschaften von Residuen zu beobachten ist. Dazu wurden GNM-Studien an den Kristallstrukturen des Ribosoms von *E. coli* und *T. thermophilus* durchgeführt. Für eine Vergleichbarkeit zwischen den GNM-Berechnungen der beiden Organismen, wurden paarweise Alignments der jeweiligen Molekülsequenzen erstellt, wenn das jeweilige Molekül in beiden Organismen vorhanden war. Da die paarweisen Alignments mit Hilfe von `clustalw` [130, 28, 78] mit Standardparametern nicht immer eindeutig gelöst wurden, sind alle paarweisen Alignments zusätzlich mit `Jalview` [31, 134] überarbeitet worden. Durch das manuelle Ändern einzelner Alignments wurde die Qualität verbessert, was in einigen Fällen zu einer Änderung in den Alignment-*Scores* geführt hat. Exemplarisch ist hier ein Ausschnitt für die Änderung innerhalb der ersten 36 Residuen des Alignments für das ribosomale Protein L29 (Abbildung 30) gezeigt.

Aus den so überarbeiteten Alignments wurde eine Übersetzungstabelle der Residuennummern von *E. coli* und *T. thermophilus* erstellt. Durch diese Übersetzungstabelle ist es möglich, die Ergebnisse der GNM-Berechnungen über die FN (Gleichung 24) zwischen der Wildtyp-Matrix und der mutierten Matrix für abgeschaltete Kontakte beider Organismen miteinander zu vergleichen. In ENM-Analysen werden unterschiedliche Distanzparameter zum Erstellen der Kontaktmatrix verwendet [83, 63]. Um die optimale Distanz zu ermitteln, die als R_c bei der Erstellung

```
E. coli          MKAKELRE--KSVEELN-TELLNLLREQ----FNLRMQAASGQ
T. thermophilus  MKLSEVRKQLEEARKLSPVELEKLVREKKRELMELRFQASIGQ

E. coli          MKAK-------ELREKSVEELNTELLNLLREQFNLRMQAASGQ
T. thermophilus  MKLSEVRKQLEEARKLSPVELEKLVREKKRELMELRFQASIGQ
```

Abbildung 30: Ausschnitt der ersten 36 Residuen des paarweisen Alignments für das ribosomale Protein L29. Überprüfung der paarweisen Alignments von *E. coli* und *T. thermophilus*. Vor (Oben) und nach (Unten) der Überarbeitung mit `Jalview` [31, 134]. Konservierte und Gap-gepaarte Residuen sind fett hervorgehoben.

der Kontaktmatrix dient, wurden drei unterschiedliche Distanzen (4,5 Å [64, 83], 6,75 Å [63] und 13 Å) genutzt. Der in ENM-Analysen häufig verwendete Schwellenwert von 13 Å, wird verwendet um nicht nur die unmittelbaren Interaktionen, sondern auch die Interaktionen mit den übernächsten Nachbarn zu berücksichtigen. Ein weiterer wichtiger Unterschied besteht in der Berechnung des Abstandes bei 13 Å, hier werden nur die C_α-C_α, P-P und C_α-P-Abstände berücksichtigt, bei den beiden anderen Schwellenwerten 4,5 Å und 6,75 Å werden jedoch die Abstände aller schweren Atome (C, N, O, P, S) gemessen. Um zu verifizieren, ob der Abstand von 13 Å auch im Ribosom die übernächsten Nachbarn abbildet, wurden Histogramme der Verteilung der Abstände innerhalb der Proteine (Abbildung 31 (A)), innerhalb der rRNA (Abbildung 31 (B)) und zwischen rRNA und Proteinen erstellt (Abbildung 31 (C)).

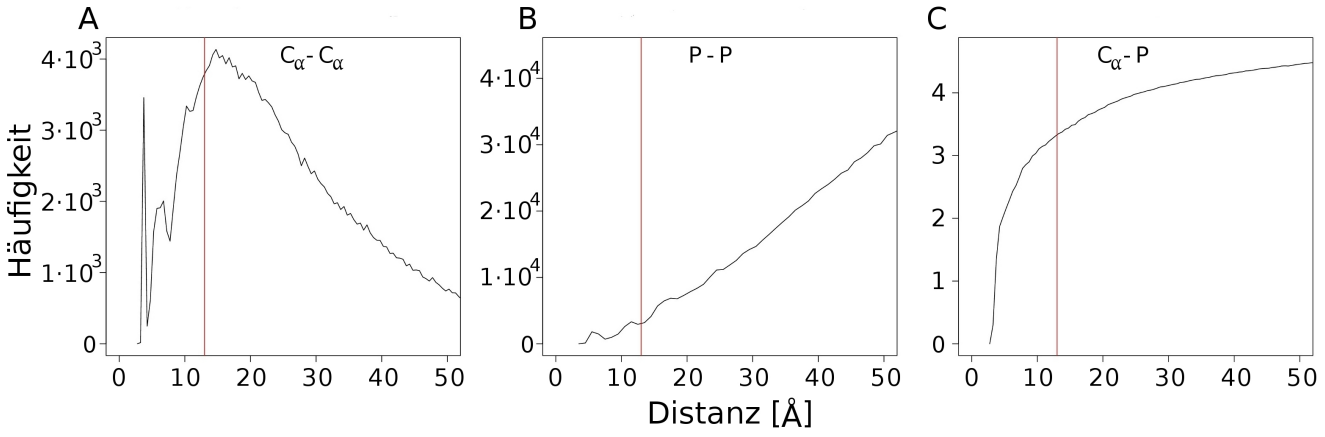

Abbildung 31: Histogramm der absoluten Häufigkeiten über alle C_α-C_α-Abstände innerhalb der Proteinketten (A), innerhalb der Phosphat-Atome innerhalb der rRNA (B) und zwischen C_α der Proteine und der Phosphat-Atom der rRNA (C). Die Häufigkeiten in (C) sind logarithmisch aufgetragen. In rot ist die Distanz von 13 Å markiert.

Abbildung 31 zeigt, dass der Abstand von 13 Å innerhalb der einzelnen Molekülklassen die Interaktionen mit den übernächsten Nachbarn sehr gut abbildet. Die Distanzverteilung in Abbildung 31 (C) ist weniger eindeutig, da sich die rRNA über das gesamte Ribosom erstreckt und so-

mit für jedes C_α-Atom eine Interaktion mit einem P-Atom zu finden ist, die in beliebiger Distanz in den räumlichen Grenzen des Ribosoms existiert. Aus diesem Grund wurde an einigen ausgewählten Stellen des Ribosoms eine visuelle Verifikation durchgeführt [Daten nicht gezeigt], aus der hervorging, dass der Schwellenwert von 13 Å für den Abstand zwischen Phosphor- und C_α-Atomen als Näherung verwendet werden kann, da die übernächsten C_α-Atome des Proteins noch mit im 13 Å-Radius um ein P-Atom der rRNA in Nähe eines gebundenen Proteins zu finden waren.

Um die Auswirkung der unterschiedlichen Distanzschwellenwerte auf die Berechnung der Kontaktmatrix zu untersuchen, wurde die GNM-Analyse (nach Kapitel 5.2.1) für alle drei Kontaktmatrizen durchgeführt und die FN-Werte gegenübergestellt (siehe Abbildung 32). Die Scatter-Plots zeigt dabei nur FN-Werte, die in der Schnittmenge der Kontakte für die jeweils verglichenen Distanzen existieren. Aus der Korrelation der FN-Werte lässt sich dann ablesen wie groß der Einfluss der unterschiedlich erstellten Kontaktmatrizen ist.

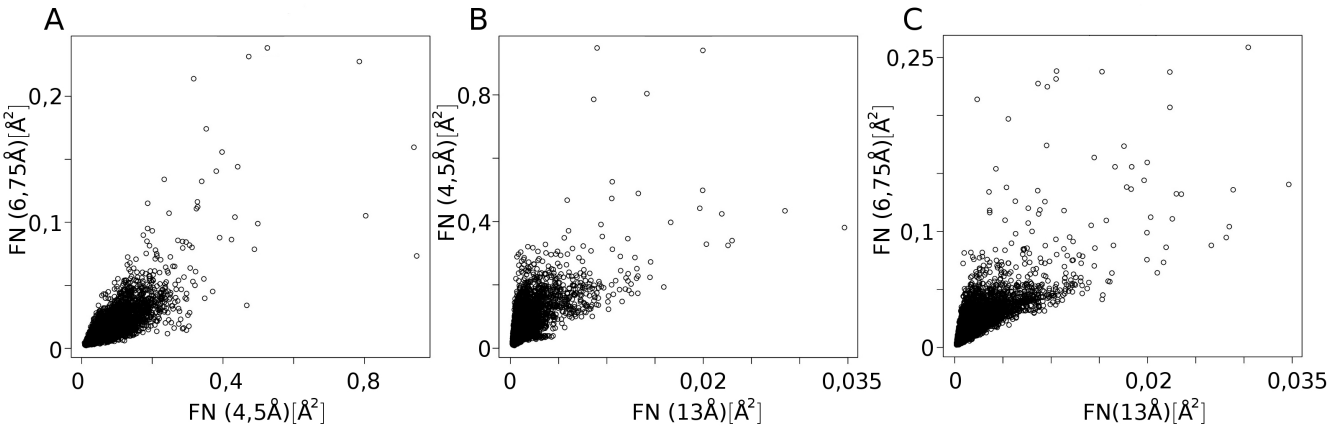

Abbildung 32: Vergleich der FN-Werte für die jeweilige Schnittmenge der Kontakte der jeweils verglichenen Distanzschwellenwerte. (A) zeigt den Vergleich von 6,75 Å zu 4,5 Å, (B) den Vergleich von 4,5 Å zu 13 Å und (C) zeigt den Vergleich von 6,75 Å zu 13 Å.

Die Vergleiche in Abbildung 32 zeigen, dass sich für alle GNM-Analysen eine positive Korrelation ergibt (Tabelle 11). Dabei sind die Korrelationen der FN-Werte für die jeweils ähnlichen Distanzen (4,5 Å & 6,75 Å bzw. 6,75 Å & 13 Å) am höchsten. Für die Ergebnisse der FN-Werte aus den GNM-Analysen mit dem größten Distanzunterschied sind entsprechend der Spearman- und der Pearson-Korrelationskoeffizient am kleinsten.

Die Tabelle 11 zeigt für beide Koeffizienten einen Zusammenhang in der Relevanz der Interaktion der einzelnen Kontakte im gleichen Größenbereich, woraus sich tendenziell eine Unabhängigkeit des Distanzschwellenwertes bei der Berechnung der Kontaktrelevanz ableiten lässt. Hierbei ist allerdings zu berücksichtigen, dass die Korrelation der FN-Werte abhängig von den verglichenen Distanzschwellenwerten abnimmt.

Auch die Skala der FN-Werte nimmt mit steigendem Distanzschwellenwert ab (vergleiche Abbildung 32). Diese Beobachtung ist durch den stabilisierenden Effekte der nächsten Nachbarn erklärt, den Maslov et al. in Protein-Protein-Netzwerken zeigen konnten [93]. Dieser Effekt

Tabelle 11: Korrelationskoeffizienten der FN-Werte für die verschiedenen Entfernungsschwellenwerte R_c für die Untersuchung der großen ribosomalen Untereinheit von *E. coli*. Im oberen Dreieck sind die Spearman-Koeffizienten der einzelnen Vergleiche und im unteren Dreieck die Pearson-Koeffizienten notiert.

Pearson \ Spearman	4,5 Å	6,75 Å	13 Å
4,5 Å		0,808	0,7
6,75 Å	0,813		0,824
13 Å	0,698	0,789	

und die Tatsache, dass in der hier durchgeführten Analyse immer nur ein einzelner Kontakt abgeschaltet wird, tragen dazu bei, dass das Ergebnis der GNM-Analyse bei einem Distanzschwellenwert von 13 Å einen zu geringen Einfluss des abgeschalteten Kontakt zeigen. Der Distanzschwellenwert von 4,5 Å liegt sehr nah an der Auflösungsgrenze von ca. 3 Å in der PDB-Struktur, sodass die GNM-Analysen mit einem Distanzschwellenwert von 6,75 Å durchgeführt werden.

Für den Vergleich der GNM-Analyse der beiden unterschiedlichen Organismen (*E. coli* und *T. thermophilus*) wurden die oben beschriebenen Übersetzungstabellen verwendet, um die Schnittmenge der Kontakte zu bestimmen. Anschließend wurden die FN-Werte der Kontaktschnittmenge wieder in einem Scatter-Plot miteinander verglichen. Die Ergebnisse dieses Vergleichs sind in Abbildung 33 (B) gezeigt. Anders als erwartet zeigen die Molekülkontakte keine ausgeprägte Korrelation zwischen den Organismen, obwohl das Ribosom in beiden Organismen die selbe Funktion aufweist. Bei der Betrachtung des Korrelationsplots sind Ausreißer zu erkennen, die eine wesentlich höhere FN in einem der beiden Organismen aufweisen, als der übersetzte Kontakt im jeweils anderen Organismus. Um eine Klassifizierung durchzuführen, wurden die FN-Wertepaare der beiden Organismen in drei Klassen eingeteilt. Für die Klassifikation wurden die FN-Wertepaare aus dem organismischen Vergleich als Endpunkte von Vektoren interpretiert und die Verteilung der Winkel analysiert (vergleiche Abbildung 33 (A)). Das Mittel der gefitteten Verteilung wurde als Hauptwinkel definiert.

Abbildung 33 (B) zeigt die FN-Wertepaare aus den beiden Organismen *E. coli* und *T. thermophilus*. Die Klassifizierung der FN-Wertepaare ergibt so eine Unterscheidung in drei Klassen. Ist das Wertepaar näher zum Hauptwinkel als zu einer der beiden Hauptachsen, wird der Einfluss auf beide Organismen als ähnlich definiert (schwarz) und der Klasse 3 (C3) zugeordnet. Ist der Abstand zur Abszisse kleiner als zum Hauptwinkel, der FN-Wert des abgeschalteten Kontakts in *E. coli* also größer als in *T. thermophilus* (grün) so wird der Kontakt der Klasse 2 (C2) zugeordnet. Zeigt das Kontaktpaar einen höheren Eintrag für *T. thermophilus* (rot), ist also der Ordinate näher als dem Hauptwinkel, so wird der Kontakt der Klasse 1 (C1) zugeordnet. Tabelle 12 zeigt eine Übersicht über die Korrelationskoeffizienten zwischen den FN-Werten von *E. coli* und *T. thermophilus* innerhalb der drei Klassen sowie der gesamten Daten.

Die Klassifizierung der FN-Wertepaare zeigt deutlich, dass in beiden Molekülen der Hauptteil der Kontakte einen ähnlich starken Einfluss auf die Mechanik des Ribosoms hat, da der

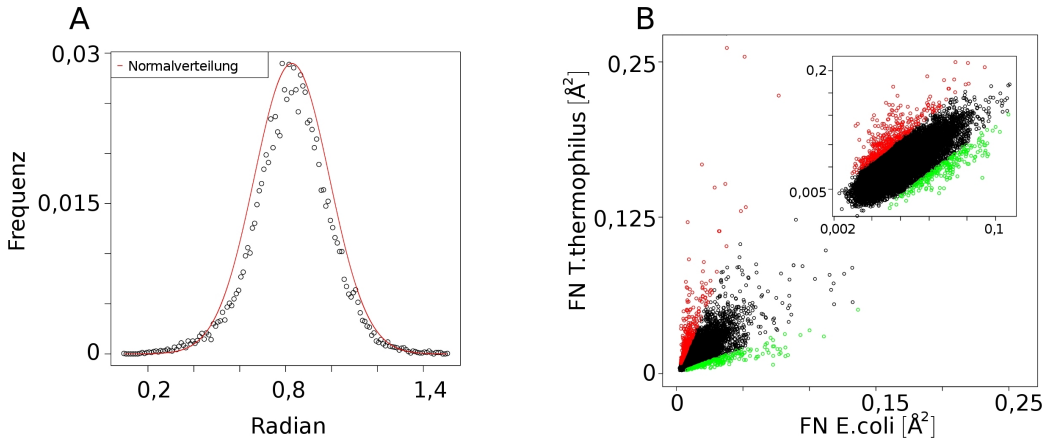

Abbildung 33: (A) zeigt die Näherung der Winkelverteilung mit Hilfe einer Normalverteilung (rot). Die Parameter der Normalverteilung wurden aus der beobachteten Verteilung geschätzt. (B) FN zwischen den Korrelationsmatrizen des Wildtyps und der Korrelationsmatrizen mit jeweils einem abgeschalteten Kontakt. Für jeden Kontakt, der sowohl in *E. coli* als auch in *T. thermophilus* besteht ist die jeweilige FN von *E. coli* gegen *T. thermophilus* aufgetragen. Farbig markiert sind Kontakte, die in beiden Organismen einen ähnlichen Einfluss (schwarz), mehr Einfluss in *E. coli* (grün) oder mehr Einfluss in *T. thermophilus* (rot) haben. Das Inset zeigt die selben Daten in einem log-log-Plots.

Tabelle 12: Korrelationskoeffizienten der Klassifizierung durch einfache Geometrie. In Klammern sind die Organismen vermerkt, auf die die Kontakte mehr Einfluss haben.

	C1 (*T. thermophilus*)	C2 (*E. coli*)	C3 (beide)	Ohne Klassifizierung
Pearson	0,798	0,905	0,817	0,686
% der Kontakte	1,5	0,8	97,7	100

Pearson-Korrelationskoeffizient bei isolierter Betrachtung der Klasse 3 von $\sim$0,7 auf 0,9 steigt. Die Abweichung vom erwarteten Ergebnis wird von einem kleinen Prozentsatz (ca. 2,5 %) der übersetzten Kontakte hervorgerufen.

Dieser kleine Prozentsatz wurde in `VMD` [71] analysiert, um über die Lage der Kontakte Rückschlüsse auf die Funktion zu ziehen. Es wird deutlich, dass die Kontakte im Ribosom verteilt vorliegen und keine spezifische Häufung in funktional wichtigen Domänen des Ribosoms zu finden sind. Genauer betrachtet sind die Kontakte der Klassen C1 und C2 in Clustern gehäuft, das heißt, dass ein Residuum mehrere Kontakte aufweist (vergl. Abbildung 34).

Trennt man die Kontakte nach inter- und intramolekular, fällt auf, dass in den Klassen C1 und C2 nur ein sehr kleiner Anteil an intermolekularen Kontakten zu finden ist. Nur jeweils ca. 3 % der Kontakte aus C1 bzw. C2 sind in der großen Untereinheit bei *E. coli* und *T. thermophilus* intermolekular. In der kleinen Untereinheit sind es ca. 9 % bei *T. thermophilus* und ca. 15 % bei *E. coli*. Die organismusspezifischen Kontakte sind also eher intramolekular als intermolekular.

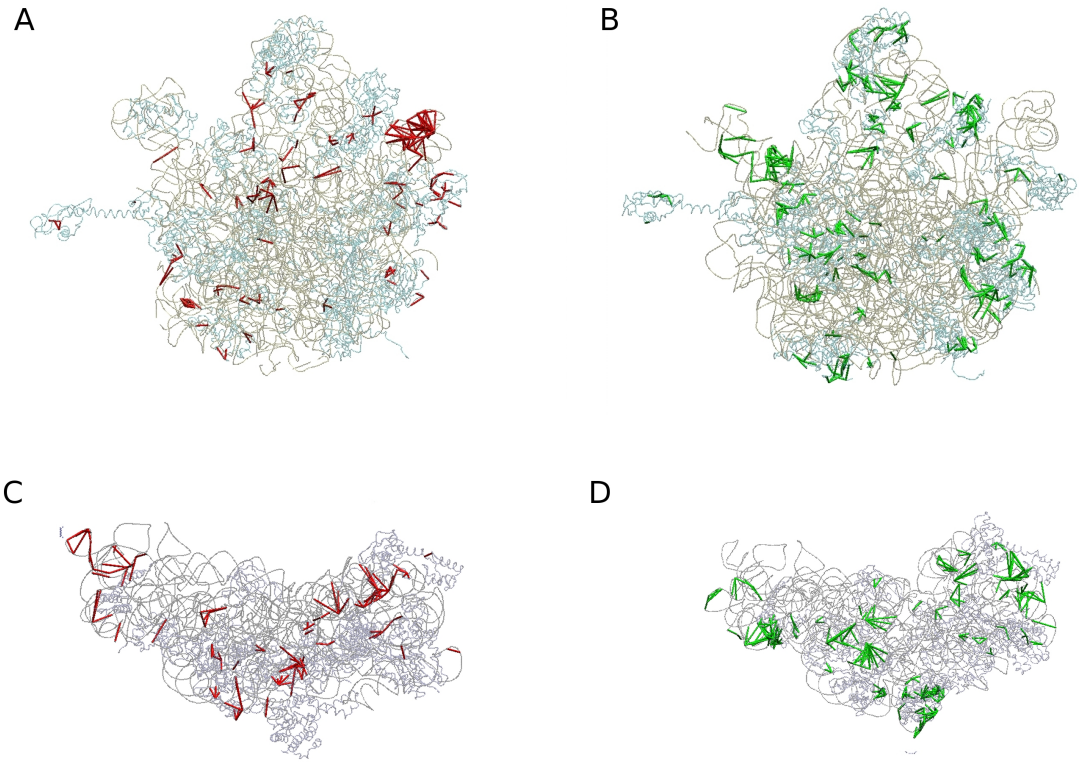

Abbildung 34: Visualisierung der Kontakt-Klassen C1 (rot) und C2 (grün) im PDB des jeweiligen Organismus. In (A) ist die Kontakt-Klasse C1 im PDB von *T. thermophilus* dargestellt. Die Kontakt-Klasse C2 ist im PDB von *E. coli* dargestellt (B). In (C) und (D) sind jeweils die Kontakt-Klassen der kleinen ribosomalen Untereinheit gezeigt.

Die organismusspezifische Evolution des Ribosoms scheint eher durch Veränderung innerhalb einzelner Molekülketten stattzufinden, als durch Veränderung der Molekülketten-Interaktionen.

5.3.1.1 Molekulare Dynamik und Koevolution

Die GNM-Analyse der beiden ribosomalen Untereinheiten zeigt bei wenigen Kontakten einen unterschiedlich starken Einfluss auf die Mechanik des Moleküls bei beiden untersuchten Organismen (Kapitel 5.3.1). Aus der Lokalisation der Residuen innerhalb der PDB-Strukturen lässt sich kein funktioneller Zusammenhang identifizieren. Um den vermuteten evolutionären Zusammenhang zu untersuchen, wurden die Ergebnisse der MI-Berechnung und die FN-Werte der GNM-Analyse miteinander verglichen. Da aus Studien der Acetylcholinesterase [137] und der HIV-1 Protease [62] bekannt ist, dass strukturell relevante Residuen auch ein hohes Maß an Koevolution aufweisen können, ist anzunehmen, dass sich auch im Ribosom Residuenpaare zu finden sind, welche sowohl eine hohe MI aufweisen, als auch einen hohen FN-Wert nach der GNM-Analyse bei ausgeschaltetem Kontakt des betrachteten Residuenpaars. Bei dieser Analyse wurden nur Residuenpaarungen berücksichtigt, für die ein FN-Wert > 0 berechnet wurde, die also einen Einfluss auf die molekulare Dynamik des Ribosoms haben. Außerdem wurde die Analyse getrennt für Inter- und Intra-Kontakte durchgeführt. Für den Vergleich wurden die GNM-

Resultate der ribosomalen Untereinheiten aus den PDBs von *E. coli* und *T. thermophilus* (Kapitel 5.3.1) mit den Z-Scores der MI-Analysen (Kapitel 3.3.1) verglichen. Für diese Untersuchung wurden Übersetzungstabellen angefertigt, die es ermöglichen, einen MI-Index in der jeweiligen PDB-Struktur wiederzufinden und damit auch auf die Ergebnisse der GNMs zu übertragen. Dazu wurden die PDB-Sequenzen an die bestehenden Alignments der mit HMM gefundenen Sequenzen für die MI-Berechnung mit Hilfe der Profilalignment-Funktion von `clustalw` [130, 28, 78] aligniert.

Für den Vergleich wurden zunächst die FN-Matrizen von ihrer globalen Form des gesamten Ribosoms in die unterschiedlichen lokalen Matrizen (Protein-Protein und Protein-rRNA) aufgeteilt. Zudem wurden anschließend aus den MI-Matrizen nur die Zeilen und Spalten verwendet, die eine Entsprechung in der jeweiligen FN-Matrix aufweisen. Somit wurden jeweils Submatrizen aus den globalen und lokalen Matrizen gebildet, die nur noch Werte der Residuen enthalten für die eine jeweilige Entsprechung in der MI-Matrix und in der FN-Matrix existiert. Aus diesen unterschiedlichen Aufteilungen der Matrizen, ergeben sich vier Möglichkeiten der Normierungen:

1. Normierung auf das globale Maximum der Submatrizen
   (das Maximum aller berechneten Werte der tatsächlich gefundenen Kontakte)

2. Normierung auf das globale Maximum
   (das Maximum aller berechneten Werte)

3. Normierung auf jeweils das lokale Maximum der Submatrizen
   (das Maximum der tatsächlich gefundenen Kontakte im jeweiligen Molkülvergleich)

4. Normierung auf das lokale Maximum
   (das Maximum aus dem jeweiligen Molkülvergleich)

Abbildung 35 zeigt die Auftragung der normierten MI gegen die normierten FN-Werte in Scatter-Plots für alle Intra-Kontakte des Ribosoms von *E. coli*.

Bei der Betrachtung der Abbildung 35 zeigt sich nur bei der Normierung beider Wertebereiche (GNM und MI) auf das lokale Maximum der jeweiligen Submatrizen Kontaktpaare, die sowohl eine hohe FN-Wert aufweisen, als auch ein hohes Signal der MI (Abbildung 35 (A)3). Alle anderen Plots zeigen keine Kontaktpaare im oberen rechten Quadranten der Scatter-Plots. Da sich nur in einem der Normierungsverfahren Kontaktpaare finden lassen, die hier als interessant definiert wurden, wird die Normierung im folgenden näher betrachtet. Dazu ist es notwendig die Ergebnisse der GNM-Analyse getrennt von den Ergebnissen der MI-Analyse zu betrachten. Die Ergebnisse der GNM-Analyse sind nur im Zusammenhang der gesamten Struktur der ribosomalen Untereinheit zu interpretieren, da sie aus Berechnungen auf Grundlage der Kontaktmatrix stammen. Daraus folgt für die Normierungen auf die Submatrizen, dass diese die Ergebnisse der GNM-Analyse überschätzt. Genau so überschätzt die Normierung auf das lokale Maximum der GNM-Analyse für die einzelne Molekülkette die FN-Ergebnisse. Aus diesen Überlegungen wird deutlich, dass die FN-Werte der GNM-Analyse auf das globale Maximum der gesamten zur Verfügung stehenden Ergebnisse normiert werden muss. Für die Ergebnisse der MI-Analyse ist es wichtig auf die unterschiedlich große Symbolmengen zu verweisen, die zur Berechnung der MI zur Verfügung standen. So führen die Ergebnisse der MI-Analyse für die RNA zu einem

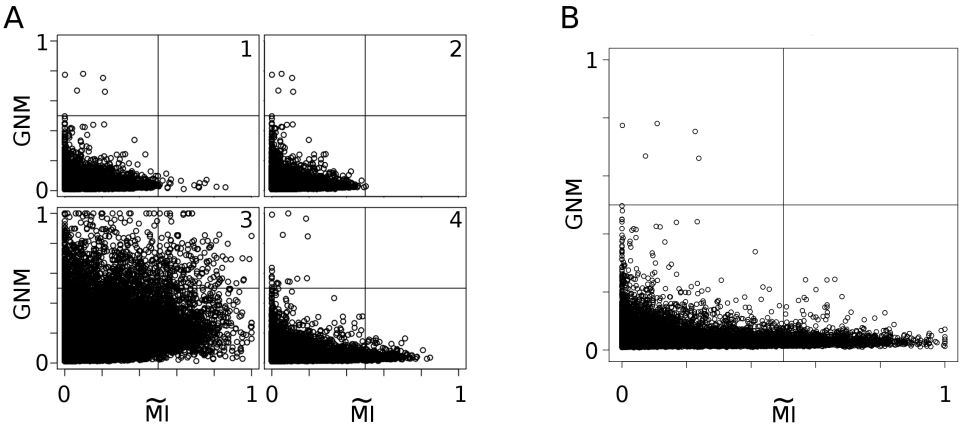

Abbildung 35: Scatter-Plots der FN-Werte aus den GNM-Analysen aufgetragen gegen die MI-Werte aller Intra-Kontakte. In (A) sind die vier verschiedenen Normierungen gezeigt. Dabei ist die Normierung auf das globale Maximum (2) und das lokale Maximum (4), sowie die Normierung auf das globale Maximum der Submatrizen (1) und das lokale Maximum der Submatrizen (3) gezeigt. In (B) wurden die Werte der GNM-Analyse auf das globale Maximum der Submatrix normiert, die MI-Werte wurden auf das lokale Maximum der Submatrizen normiert.

anderen theoretischen Maximum, durch die Abschätzung nach Gleichung 5, als die Ergebnisse der MI-Analyse der Proteine. Daraus ergibt sich eine Unterschätzung einzelner MI-Ergebnisse durch die Normierung auf das globale Maximum. Aus dieser Betrachtung wird die Verwendung von unterschiedlichen Normierungen für beide Analysen ersichtlich, nämlich die Normierung der MI auf das lokale Maximum der Submatrizen und die Normierung der GNM-Analyse auf das globale Maximum. Wie Abbildung 35 (B) zeigt, existiert bei diesem Normierungsverfahren allerdings kein FN-MI-Wertepaar im oberen rechten Quadranten, was darauf schließen lässt, dass sich in der großen ribosomalen Untereinheit kein strukturell wichtiger Kontakt durch eine hohe koevolutionäre Interaktion auszeichnet. Aus diesen Ergebnissen lässt sich kein Einfluss von koevolutionären Signalen auf die molekulare Dynamik feststellen. Die GNM-Analyse der PDB-Strukturen aus verschiedenen Organismen zeigt, dass es sich bei diesen Ergebnissen eher um eine generelle Aussage handelt, als um eine organismusspezifische Beobachtung. Dies bestätigt die Beobachtungen von Hamacher [62] und Osadchy et al. [103], dass die Faltung einen höheren Einfluss auf die Molekül-Funktionalität hat als einzelne Mutationen der Primärsequenz.

5.3.1.2 Evolutionsmatrizen

In der anschließenden Betrachtung soll das Verhältnis zwischen Koevolutionssignal und dem Signalen der GNM-Analysen untersucht werden. Dazu wurden die Ergebnisse der zwei unterschiedlichen Analysen (GNM und MI) jeweils auf den Wertebereich zwischen Null und eins normiert. Zu diesem Zweck wurden die Ergebnisse der GNM-Analyse und der MI-Analyse jeweils in Evolutionsmatrizen (Kapitel 3.2.6) überführt, um so eine vergleichbare Besetzung der Matrizen zu erhalten. Dabei wurden die beiden Untereinheiten des Ribosoms getrennt voneinander untersucht, da die GNM-Analysen nur für die einzelnen PDB-Strukturen durchgeführt

wurden. Somit werden die Interaktionen zwischen der großen und kleinen Untereinheit nicht berücksichtigt.

Da die GNM-Analyse die Wichtigkeit eines Kontakts analysiert, werden viele Molekül-Paarungen, nämlich genau solche, die nicht in physischem Kontakt stehen mit Null besetzt. Die daraus entstehenden Evolutionsmatrizen sind somit mit vielen Nullen besetzt. Um eine Vergleichbarkeit zwischen den Evolutionsmatrizen aus den GNM-Analysen mit den Evolutionsmatrizen der *Z*-Score-Matrizen zu erreichen, wurde die Dichte der Evolutionsmatrizen aus den GNM-Analysen bestimmt, in dem der Anteil an Werten ungleich Null ermittelt wurde. Aus den *Z*-Score-Evolutionsmatrizen wurden dann der gleiche Anteil der größten Werte für den anschließenden Vergleich verwendet. Die vier resultierenden Evolutionsmatrizen der großen ribosomalen Untereinheit von *T. thermophilus* (PDB-Code 2AWB) sind in Abbildung 36 gezeigt. Die Dendrogramme in Abbildung 36 zeigen keine vergleichbaren Cluster-Strukturen zwischen den Evolutionsmatrizen der GNM- und MI-Analysen. Daraus lässt sich schließen, dass auch auf globaler Ebene der Evolutionsmatrizen kein Zusammenhang zwischen dem koevolutionären Signal und der molekularen Dynamik im Rahmen der ENM-Näherung innerhalb der ribosomalen Untereinheiten besteht. Dieses Ergebnis unterstreicht so noch einmal die der GNM-Analyse (vergleiche Kapitel 5.3.1.1).

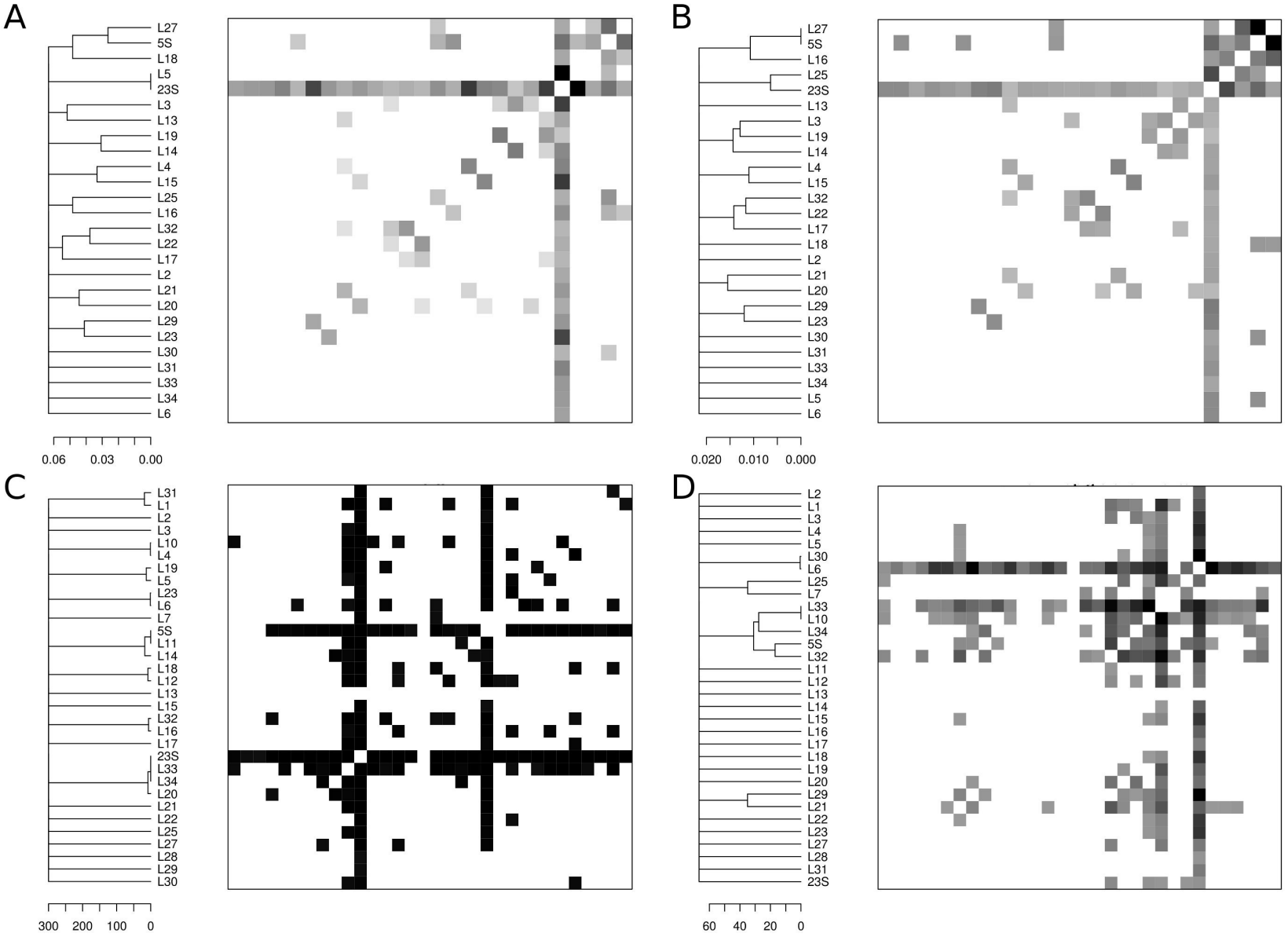

Abbildung 36: Darstellung der vier Evolutionsmatrizen aus dem Maximum der Z-Scores und der FN-Werte in (A) und (B), sowie der mittleren Z-Scores und FN-Werte in (C) und (D). Links neben den Matrizen sind jeweils die Dendrogramme der `hclust`-Ergebnisse gezeigt. (B) und (D) in $Å^2$

5.3.2 SCPCP

Die gewonnenen Einblicke in die Thermodynamik der kleinen Untereinheit [65] sollen hier für die große Untereinheit von *T. thermophilus* (PDB-Code 2AW4) nachvollzogen werden. Der hauptsächliche Unterschied zu der kleinen Untereinheit besteht nicht nur in der unterschiedlichen Anzahl an ribosomalen Proteinen, sondern auch darin, dass in der großen Untereinheit eine zusätzliche ribosomale RNA, die 5 S rRNA, enthalten ist, die in der Berechnung zu berücksichtigen ist. In der hier gezeigten Untersuchung wurde die 5 S rRNA wie ein ribosomales Protein im Protokoll von Hamacher et al. [65] behandelt und aus dem Komplex entfernt. Aus den folgenden SCPCP-Berechnungen wurde dann durch die Bestimmung der thermodynamischen Unterschiede in der Bindungsenergie, sich beeinflussenden Molekülketten bestimmt (Abbildung 39). Die Matrix zeigt auf der „x-Achse“ die Molekülketten, die die Bindung der Molekülketten auf der „y-Achse“ beeinflussen. In der Graphik ist die Stärke der Beeinflussung ausgedrückt durch die Anzahl an Rängen, um die die Bindungsenergie abnimmt.

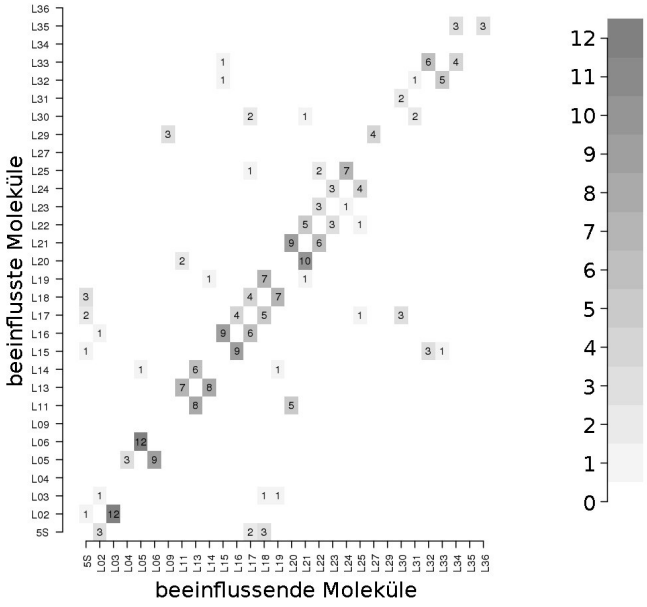

Abbildung 37: Stärke der Beeinflussung einer Molekülkette (x-Achse) auf eine andere Molekülkette (y-Achse). In der Matrix ist die Stärke durch die Änderung des Rangs notiert.

Aus dieser Matrix lässt sich mit Hilfe von `GraphViz` [43] ein Netzwerk erstellen, das die Beeinflussung der verschiedenen Moleküle der großen ribosomalen Untereinheit in zirkulärer Form darstellt (Abbildung 39 (A)). Diese Darstellungsform wurde auch für die von Herold et al. [66] publizierte Bindeabhängigkeit der einzelnen ribosomalen Proteine gewählt und neben dem berechneten Netzwerk in Abbildung 39 (B) gezeigt. Der Vergleich der beiden Netzwerke (Abbildung 39) zeigt keine gute Übereinstimmung. Vor allem die zentrale Rolle des Proteins L15 im Netzwerk von Herold et al. [66] kann durch die SCPCP-Analyse nicht reproduziert werden. Auch das Cluster der Proteine L5, L15 und L18, das die Bindung der 5 S rRNA vermittelt, wird mit der SCPCP-Analyse nicht identifiziert. Ein weiterer Unterschied zu den experimentellen Ergebnissen von Herold et al. ist die fast ausschließlich gegenseitige Beeinflussung der ribosomalen Proteine, die sich durch die SCPCP-Analyse ergibt. Das zeigt auch der fast symmetrische Aufbau der in

Abbildung 37 gezeigten Matrix, wohingegen die Pfeile in Abbildung 39 (B) zumeist nur in eine Richtung zeigen, und somit eine einseitige Beeinflussung der Bindung symbolisieren.

Um diese signifikanten Unterschiede aufzuklären, wurden die B-Faktoren der SCPCP-Berechnung mit den experimentell bestimmten B-Faktoren aus dem PDB verglichen. Dazu wurden die B-Faktoren der C_α-Atome mit den B-Faktoren der SCPCP-Analyse über den Spearman-Korrelationskoeffizienten verglichen. Die Ergebnisse des Vergleichs sind in Abbildung 38 gezeigt.

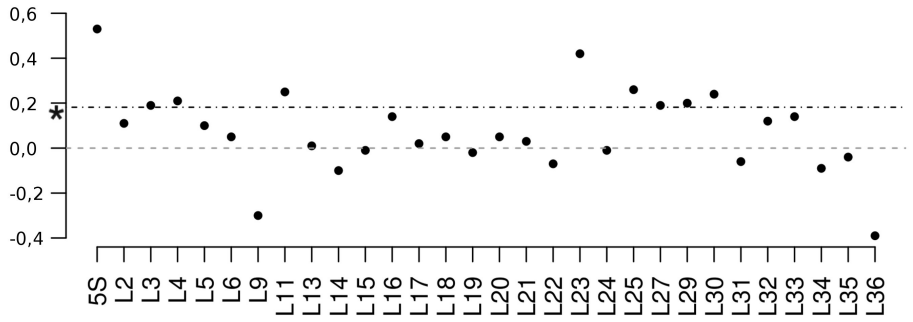

Abbildung 38: Spearman-Korrelationskoeffizienten der B-Faktoren aus der SCPCP-Berechnung für die 50 S Untereinheit des Ribosoms aus *T. thermophilus* und den B-Faktoren aus der PDB-Struktur. Zur Orientierung ist die Null-Linie in hellgrau markiert und die minimale Spearman-Korrelation aus den Untersuchungen von Hamacher et al. [65] durch eine Linie und ein (∗) markierte.

Die Übersicht der B-Faktor-Korrelation zeigt eine schlechte Übereinstimmung der B-Faktoren für die meisten Moleküle. Die von Hamacher et al. definierte minimale Korrelation von 0,182 (Abbildung 38 (∗)), wird benötigt um eine Signifikanz von 95 % [65] in der Übereinstimmung der B-Faktoren zwischen den experimentell bestimmten Werten und den berechneten aus der SCPCP-Analyse zu erreichen. Nur neun der insgesamt 30 Molekülketten liegen oberhalb dieser minimalen Korrelation.

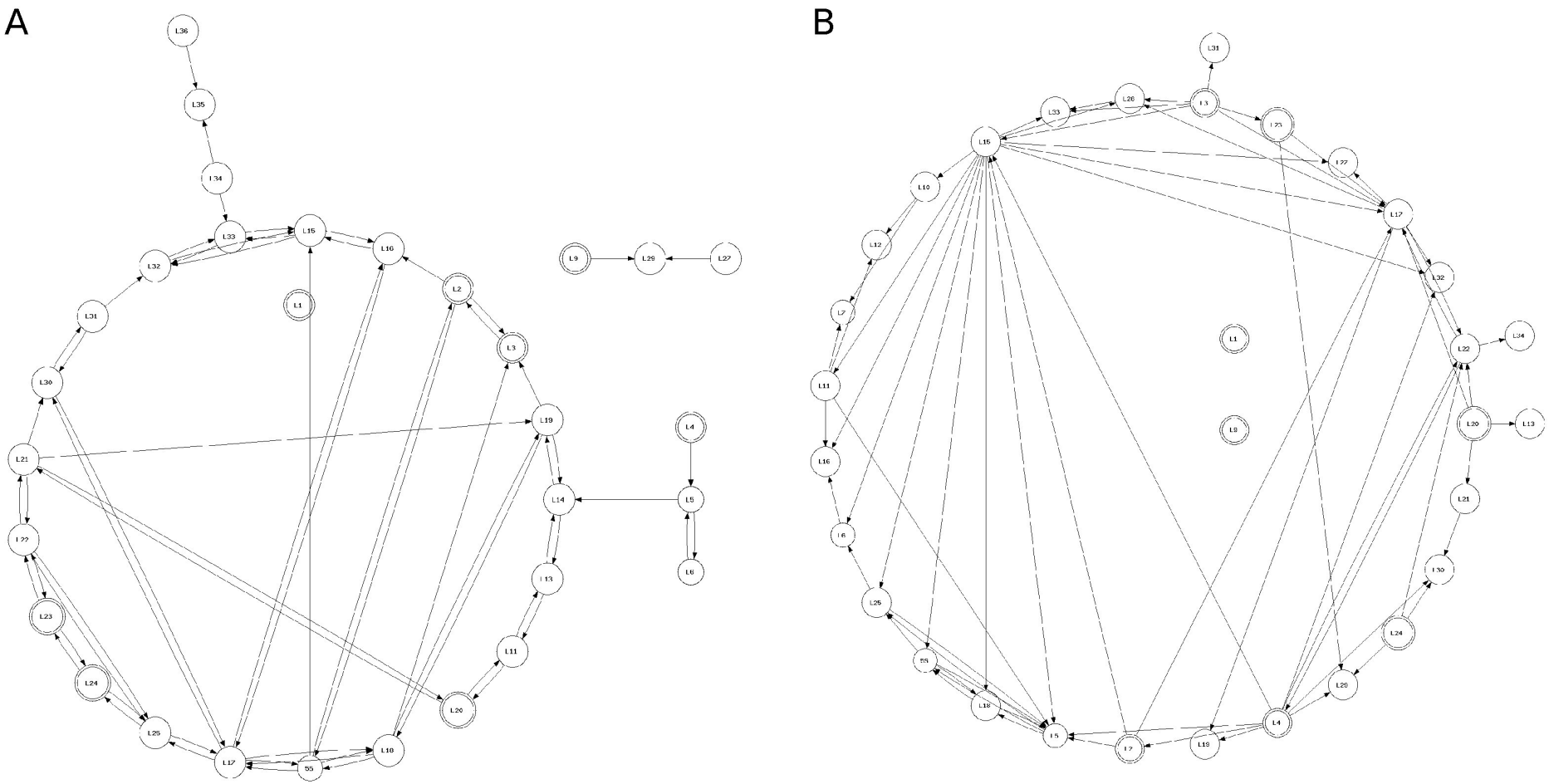

Abbildung 39: (A) Zirkuläre Darstellung der Bindung beeinflussenden ribosomalen Proteine in der großen ribosomalen Untereinheit. Die Pfeilrichtungen zeigen die beeinflussende Richtung an, L9 → L29 heißt also, dass die Präsens von L9 die Bindung von L29 verbessert. (B) Zirkuläre Darstellung der von Herold et al. [66] identifizierten Beeinflussung der Bindung. Die doppelt umkreisten Proteine sind die durch Herold et al. [66] identifizierten primär bindenden Proteine.

5.4 Diskussion

5.4.1 GNM

Es konnte gezeigt werden, dass die für die GNM-Analysen benutzten Distanzschwellenwerte zu unterschiedlichen Ergebnissen führen, dabei ist vor allem die stärke des Einflusses vom Distanzschwellenwert R_c abhängig. Mit steigendem R_c nimmt die Stärke des Einflusses, also der Höhe des resultierenden FN-Wertes, ab. Dieser Effekt ist durch den stabilisierenden Effekt höherer Vernetzung zu erklären, der von Maslov et al. [93] 2002 für Proitein-Interaktions-Netzwerke gezeigt wurde. Die Methode der Kontaktbestimmung scheint einen größeren Einfluss zu haben als die Distanz selbst (vergleiche Korrelationskoeffizienten von 4,5 Å und 6,75 Å mit dem Korrelationskoeffizienten von 13 Å Tabelle 11). Generell ist zu erwarten, dass die Distanzberechnung über die schweren Atome zu einem detaillierten Modell führt, da Ausrichtung und Struktur der einzelnen Seitenketten berücksichtigt werden. Die allgemeine Korrelation zeigt eine generelle Nähe der Ergebnisse, die sich nur im Detail unterscheidet, wie es bei solch unterschiedlichen Verfahren nicht anders zu erwartet ist. Da die Auflösungen der PDB-Strukturen, die in dieser Arbeit verwendet wurden, bei ca. 3 Å liegen, ist durch die relative Unschärfe der Struktur eine ungenaue Abbildung der strukturellen Eigenschaften der Kristallstruktur zu befürchten. Wird der Distanzschwellenwert von 4,5 Å gewählt, liegt der mögliche Fehler nur knapp unterhalb der Distanz einer Wasserstoffbrückenbindungen von 1,8 Å [133] und wurde daher nicht weiter berücksichtigt.

Die Vergleiche der Ergebnisse der beiden Organismen *T. thermophilus* und *E. coli* ist in erster Linie vom Alignment abhängig, aus dem die Übersetzungstabellen erstellt wurden. Da jeder verwendete Alignment-Algorithmus dahingehend unterschiedlich optimiert ist, das lokale Minimum der paarweisen Alignments zu finden, kann dies zum Einen zu mehrfachen möglichen Lösungen des Alignments und damit zum Anderen zu unterschiedlichen Ergebnissen des GNM-Vergleichs führen. Um dieses Problem zu lösen, haben wir einen visuellen Vergleich der unterschiedlichen Ergebnisse mit Hilfe von `JalView` [31] durchgeführt. Dabei haben wir aus den Ergebnissen des Alignment-Algorithmus `clustalw` [130, 28, 78] immer das Alignment gewählt, dass die größten zusammenhängenden Sequenzblöcke aufwies, um eine größtmögliche Anzahl an Kontakten zu übersetzen.

Die Analyse der beiden ribosomalen Untereinheiten zeigt eine geringere Korrelation der FN-Werte als erwartet (siehe Abbildung 33 (B)). Diese geringe Korrelation zeigt, dass trotz der funktionalen Ähnlichkeit der beiden Molekülkomplexe Kontakte existieren, die unterschiedliche Auswirkungen auf die Stabilität des Ribosoms in den unterschiedlichen Organismen haben. Dies könnte auf die unterschiedlichen Lebensräume des extrem thermophilen *T. thermophilus* [88] und des relativ weit verbreiteten *E. coli* zurückzuführen sein, da die abiotischen Faktoren zu verschiedenen Stabilitätskriterien innerhalb des Ribosoms geführt haben könnten. Diese Vermutung lässt sich durch den Vergleich mit der koevolutionären Signatur stützen. Die MI-Analyse beinhaltet allgemeine Aussagen zur Koevolution, weil die zu Grunde liegenden Alignments nicht auf die Sequenzen von *E. coli* und *T. thermophilus* beschränkt sind, sondern ca. 320 verschiedene Bakterienstämme enthalten. Dadurch können solch organismenspezifische Signaturen (wie hier in der GNM-Analyse beobachtet) maskiert sein.

5.4.2 SCPCP

Die SCPCP-Analyse zeigt keine Übereinstimmung mit bisher publizierten Assemblierungs-Maps [66]. Dabei hat das Verfahren der SCPCP-Analyse für die kleine ribosomale Untereinheit eine sehr gute Übereinstimmung mit der Assemblierungs-Map der kleinen Untereinheit gezeigt und ließ sogar neue Schlüsse über noch nicht näher analysierte Teile zu (Hamacher et al. [65]). Die für die große ribosomale Untereinheit hier nicht ausreichende Übereinstimmung der Assemblierungs-Map kann an der schlechten Korrelation der B-Faktoren liegen, die sich während der SCPCP-Berechnung ergaben.

Interessant ist auch die zumeist gegenseitige Beeinflussung der Molekülketten in den hier gezeigten Ergebnissen, die so in der Assemblierungs-Map von Herold et al. [66] nicht zu finden ist. Daraus ergibt sich auch die zirkuläre Form des resultierenden Netzwerkes, das sich deutlich von der hierarchischen Struktur von Herold et al. [66] unterscheidet.

Die schlecht korrelierenden B-Faktoren können durch den komplexeren Aufbau der großen ribosomalen Untereinheit stammen, in der sich zunächst intermediäre Molekülkomplexe bilden, bevor Sie an der 23 S rRNA assoziieren. Dies wäre zum Beispiel für das Cluster aus den Proteinen L5, L15 und L18, die die Bindung der 5 S rRNA vermittelt [66], eine Möglichkeit. Mit diesem Vorwissen könnte man eine veränderte Entfernung der Molekülketten ableiten und für die SCPCP-Analyse verwenden. Das würde allerdings dazu führen, das Molekülkomplexe mit ähnlichem Assemblierungsverhalten (es bilden sich zunächst Subcluster, die zum letztendlichen Molekül assemblieren) ohne Vorwissen mit Hilfe der SCPCP nicht analysiert werden können.

Es bleibt festzuhalten, dass thermodynamische Untersuchungen an der großen ribosomalen Untereinheit einer differenzierteren Analyse bedürfen, bei der unter Anderem die Klärung wie die Bindung der 5 S rRNA in dem Modell der SCPCP behandelt werden kann, näher betrachtet werden muss, um die Ergebnisse von Hamacher et al. auf die große Untereinheit zu transferieren. Außerdem sollte für die große ribosomale Untereinheit geklärt werden, wieso die B-Faktoren aus dem PDB nicht ausreichend mit den Berechneten übereinstimmen.

6 Phylogenie Software

6.1 Einleitung

Die hier vorgestellte Visual Analytics (VA) Software `ViPhy` wurde von Bremm et al. im Institut für Graphisch-Interaktive Systeme (GRIS) entwickelt. Bei dieser Entwicklung wurden die biologischen Modelle, die phylogenetischen Daten sowie bioinformatische Expertise beitragen und das Applikationsbeispiel von uns entwickelt [16].

Auf Basis von Sequenzalignments wird die Ähnlichkeit von Sequenzen bestimmt. Diese Untersuchungen geben Aufschluss über die Herkunft von Molekülen. Aber nicht nur die Herkunft einzelner Moleküle kann untersucht werden, sondern auch Verwandtschaftsgrade verschiedener Spezies werden so bestimmt. Bei dieser Bestimmung ist die 16 S rRNA die Sequenz auf deren Grundlage der Verwandtschaftsgrad bestimmt und so Phylogenien aufgebaut werden [141]. Um diese phylogenetischen Systematiken zu visualisieren, werden die Daten in Baumstrukturen dargestellt. Diese phylogenetischen Bäume sind von einem Ausgangspunkt (der Wurzel) über Verzweigungen, der evolutionären Entwicklung der Organismen, dargestellt, deren Blätter die Artnamen enthalten.

Für das Erstellen und das Analysieren dieser phylogenetischen Bäume werden verschiedene Methoden verwendet. In erster Linie stehen dafür Alignments (Kapitel 2.2.1) zur Verfügung, auf deren Grundlage phylogenetische Bäume erzeugt werden. Dazu werden unterschiedliche Algorithmen, wie zum Beispiel `clustalw` [130, 28, 78] oder der `MUSCLE`-Algorithmus [41] verwendet. Die Berechnung der Alignments zu phylogenetischen Bäumen kann unter anderen mit PhyLip [45] durchgeführt werden. Die anschließende Visualisierung kann mit Hilfe von FigTree [97] berechnet und manuell bearbeitet werden. Eine Sammlung der verschiedenen Methoden steht auf Phylogeny.fr [37] zur Verfügung. Für den Vergleich von zwei phylogenetischen Bäumen stehen bisher Programme wie TOPD/FMTS [111] zur Verfügung, die globale Unterschiede quantifizieren können. Eine detaillierte Analyse mehrerer Bäumen gleichzeitig ist mit diesen Methoden bisher nicht möglich, es bleibt zumeist nur der visuelle Vergleich der Bäume auf dem Papier. Durch die im Folgenden vorgestellte Software soll der Vergleich mehrerer phylogenetischer Bäume gleichzeitig ermöglicht werden.

6.2 Methoden

6.2.1 Phylogenie

6.2.1.1 ViPhy

Die Software stellt zur Analyse von phylogenetischen Bäumen vier verschiedene Maße zur Verfügung. Phylogenetische Bäume bestehen aus unterschiedlichen Strukturen, die in einer Analyse berücksichtigt werden müssen. Die Strukturen sind durch den Aufbau der Bäume (T) definiert, die sich aus Verzweigungen (n), Ästen (e) und Blättern (L) zusammensetzten. Für jede Struktur

wurde ein eigenes Maß implementiert. Zunächst ist das Blatt-basierte Maß zu nennen, das eine normalisierte Variante des Robinson-Foulds-Maßes [98] darstellt:

$$s(T_1, T_2) = \frac{|L(T_1) \cap L(T_2)|}{|L(T_1) \cup L(T_2)|} \tag{31}$$

Für die Globalisierung des Maßes wird über die Distanzen aller Unterbäume gemittelt. Da dieser Vergleich nur die Blätter berücksichtigt, wird die eigentliche Struktur der Bäume in diesem Maß vernachlässigt.

Um auch die Strukturen der Bäume zu vergleichen, wurde ein neues Maß eingeführt, das die Bäume elementweise vergleicht. Dieses Maß ist nach Bremm et al. [16] definiert als:

$$s(T_1, T_2) = \frac{|\text{Element}(T_1) \cap \text{Element}(T_2)|}{|\text{Element}(T_1) \cup \text{Element}(T_2)|}, \tag{32}$$

mit $\text{Element}(T_i) = \{\{L(T^n)\}, \forall n_i \in T_i\}$für $i \in \{1, 2\}$.

Die dritte Methode zur globalen Unterscheidung der Bäume ergibt sich durch die differentielle Betrachtung der Astlängen, die in der Phylogenie die evolutionäre Distanz zweier Blätter (auch Spezies) im phylogenetischen Zusammenhang des Baumes darstellt. Dieses Maß ist inspiriert durch Steel und Penny [127], die als Maß den Unterschied der verbindenden Äste zweier Blätter definierten, und wird beschrieben durch:

$$s(T_1, T_2) = \quad 1 - \frac{ed(T_1,T_2)}{max\, wd(T_1,T_2)} \tag{33}$$

$$\text{, mit } ed(T_1, T_2) = \quad \sqrt{\sum((wd(n_i^1, n_j^1) - wd(n_i^2, n_j^2))^2} \tag{34}$$

$$\text{und } max\, wd(T_1, T_2) = \quad max\,(max\, wd(n_i^1, n_j^1), max\, wd(n_i^2, n_j^2)), \tag{35}$$

mit $\forall_{n_i^1, n_j^1} \in L(T_1), n_i \neq n_j$ und $\forall_{n_i^2, n_j^2} \in L(T_2), n_i \neq n_j$. Somit trägt dieses Maß zur evolutionären Unterscheidung der Bäume bei. Dabei ist das Maß aus Gleichung 33 aber abhängig von der Baumgröße, da es bei großen Bäumen gegen 1 konvergiert, wie Steel und Penny [127] schon diskutiert haben. Außerdem ist durch die Normierung mit der maximale Astlänge nur der Vergleich ganzer Bäume möglich. Diese Einschränkung führt dazu, dass die Subbäume immer mit der maximalen Astlänge des korrespondierenden Baumes verglichen werden.

Für die Demonstration der Software wurden 34 rRNA- und Protein-Alignments des 70 S Ribosoms aus 32 unterschiedlichen Bakterienstämmen verwendet. Diese Sequenzen wurden mit Hilfe von HMM aus der Pfam-DB in Genomsequenzen aus der GB-DB extrahiert. Um die phylogenetischen Bäume zu erstellen wurden die Sequenzdaten mit den auf der Phylogeny.fr [37] bereitgestellten Algorithmen `MUSCLE` [41] aligniert, mit `GBLOCKS` [26] überprüft und mit `PhyML` [4, 58] zu phylogenetischen Bäumen verrechnet.

6.3 Resultate

6.3.1 Phylogenetische Analysen

In der globalen Analyse der phylogenetischen Bäume werden die verschiedenen Maße (Kapitel 6.2.1.1) berechnet und in einer globalen Matrix dargestellt. Diese Übersicht dient der Orientierung, in der der Benutzer einen Referenzbaum auswählen kann. In dem hier gezeigten Beispiel eignen sich die Bäume der beiden ribosomalen Proteine S14 und S16 am besten als den Referenzbaum, da sie im Bezug auf das Element-basierte Maß (Gleichung 32), ein besonders unterschiedliches globales Maß zu den meisten anderen Bäumen aufweisen, was durch die hohen Score-Summen in der globalen Matrix zu erkennen ist. Da in der phylogenetischen Betrachtung der Organismen die 16 S rRNA als allgemein anerkannter Standard definiert ist, wird auch in unserem Beispiel dieser Baum als Referenz gewählt, um zu zeigen, wie unterschiedlich stark die Phylogenien der Proteine von der 16 S rRNA abweichen können.

Um weitere interessante Bäume zu identifizieren, können die unterhalb der globalen Matrix angezeigten Histogramme der Score-Verteilungen zur Hilfe genommen werden. Besonders Bäume, deren Scores eine bimodale Verteilung aufweisen, sind interessant, da in diesen Bäumen sowohl sehr ähnliche, als auch sehr unterschiedliche Strukturen vorhanden sind. In dem hier gezeigten Beispiel fällt dabei zusätzlich zu den beiden schon in der globalen Analyse interessanten Proteinen das ribosomale Protein L18 auf, das einen sehr kleinen Score zum Referenzbaum der 16 S rRNA aufweist, bei detaillierter Analyse jedoch einzelne Strukturen des Referenzbaumes nicht hat. Um solche Cluster über alle vorhandenen Bäume zu identifizieren, können mit Hilfe der Software alle Bäume geöffnet und Subcluster farblich hervorgehoben werden. Zusätzlich können die Bäume kollabiert werden, sodass nur besonders unähnliche Strukturen sichtbar bleiben. So werden die Unterschiede, in Abhängigkeit vom ausgewählten Score, in den verschiedenen Bäume sichtbar.

Die beschriebenen Vergleiche sind in folgender Abbildung 40 zusammengefasst:

Mit der hier vorgestellten Software `ViPhy` [16] ist es möglich, die im Folgenden aufgeführten Mustererkennung visuell und benutzergesteuert durchzuführen:

- Identifikation von global interessanten Bäumen zur Festlegung des Referenzbaumes
- Identifikation von konservierten Subclustern in allen Bäumen
- Verteilung von Subclustern in allen Bäumen
- Identifikation von globalen und lokalen Ähnlichkeiten bzw. Unterschieden
- Identifikation von allen (Un)ähnlichkeiten innerhalb aller Bäume im Vergleich mit dem Referenzbaum

Unser Beispiel zeigt das Potential der entwickelten Software, mit der es möglich ist, interessante Strukturen in mehreren zu untersuchenden phylogenetischen Bäumen einfach zu visualisieren. Weitere interessante Beispiele für die Anwendung der Software können im Bereich der Identifi-

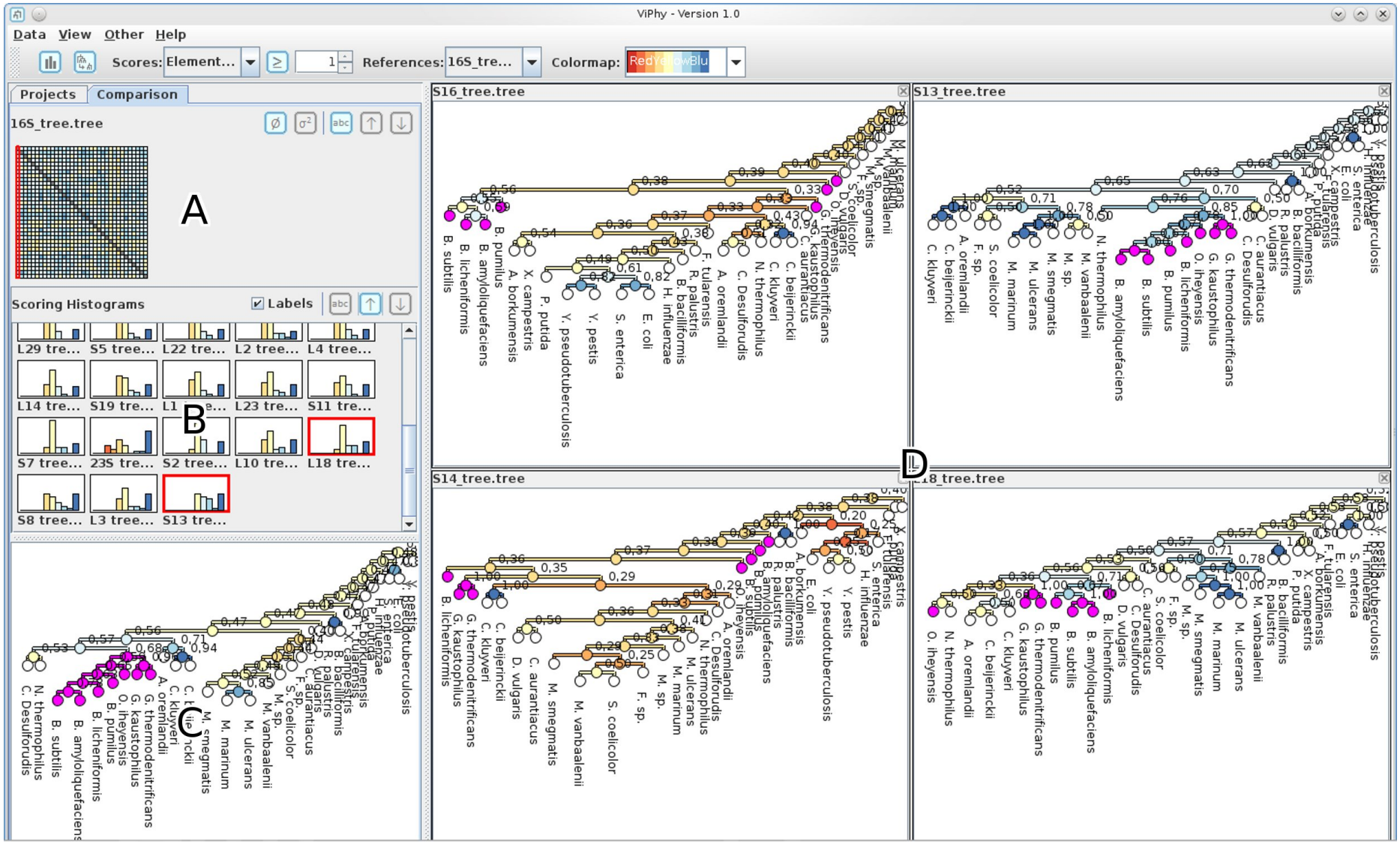

Abbildung 40: Aufbau des `ViPhy`-Fensters mit der globalen Score-Matrix (A), den Histogrammen der Score-Verteilungen (B) zum Referenzbaum (C) und der visuellen Darstellung vier einzeln ausgewählter Bäume (D). In magenta markiert ist ein Referenz-Cluster, dass in den ausgewählten Bäumen auch hervorgehoben wird.

kation von Proteinen gefunden werden, die durch horizontalen Gentransfer in unterschiedlichen Organismen zu finden sind. So können mit Hilfe von `ViPhy` [16] die phylogenetischen Abstände in Alignments vieler Proteine parallel analysiert werden und so Organismen identifiziert werden, die mehrere Proteine ausgetauscht haben. Zusätzlich können Bootstrapping-Methoden überprüft werden, die zur Verifizierung von Phylogenien herangezogen werden. So könnten zum Beispiel Spezies, deren phylogenetische Einordnung noch unklar ist und in unterschiedlichen Bootstrapping-Methoden in verschiedenen Clustern eingeordnet werden, schnell identifiziert und analysiert werden.

6.4 Diskussion

Die im Kapitel 6.3.1 gezeigten Ergebnisse zeigen keinen klassischen Ansatz phylogenetischer Untersuchungen, da die Phylogenie der Proteine eine andere als die der 16 S rRNA ist, auf Basis derer die phylogenetische Verwandtschaft zwischen Organismen aufgeklärt wird. Dennoch wird an diesem Beispiel deutlich welchen Vorteil diese Software gegenüber der bisherigen Analyse-Tools hat.

So ist die Identifikation des ribosomalen Proteins S14 interessant, weil hier eine größere Ähnlichkeit der Phylogeniebäume zwischen S14 und 16 S rRNA erwartet wurde. Da das S14 Protein über eine hoch konservierte Funktion in der Assemblierung des Ribosoms verfügt [70], ist eine konservierte Sequenz zu erwarten, die zu einem dem 16 S rRNA ähnlichen phylogenetischen Baum führen sollte. Die Beobachtung, dass sich die phylogenetischen Bäume der Proteine sehr stark unterscheiden, und die Beobachtung aus Kapitel 5.3.1.1, dass die Koevolution der Residuen nicht mit einem hohen strukturellen Einfluss eines Kontakts zusammenhängt, zeigt erneut, dass die Sequenz nicht mit der Funktion korrelieren muss.

Mit der hier vorgestellten Software `ViPhy` ist es möglich phylogenetische Bäume auf mehreren Detailstufen zu analysieren und miteinander zu vergleichen. Der Vergleich beruht hier aber in erster Linie auf, einem in der Biologie üblichen Analyse-Szenario, das aus *rooted* Bäumen mit identischen Blättern, also identischen Spezies, besteht. Für Bäume mit unterschiedlichen Blättern müssten die Maße angepasst werden. Die Skalierbarkeit dieses Analyseverfahrens liegt in der Interaktion des Benutzers, der über die verschiedenen Detailstufen (von der globalen Matrix hin zum einzeln visualisierten Baum) die Möglichkeit hat, auch eine große Anzahl von phylogenetischen Bäumen miteinander zu vergleichen. Hier ist allerdings die Darstellungsgröße durch die physikalische Auflösung ein limitierender Faktor um die einzelnen Bäume noch miteinander vergleichen zu können.

7 Fazit

Unter molekularer Koevolution versteht man die sich in der Primärsequenz biologischer Molekülen fixierenden Mutationen. Kausal liegen dieser koevolution verschiedene Selektionsfaktoren zugrunde, wie zum Beispiel physikalische Interaktionen der koevolvierenden Residuen oder der Selektionsdruck durch biochemische Interaktionen, räumlichen Nähe oder durch Liganden induzierte Wechselwirkungen.

Die Koevolution innerhalb biologischer Sequenzen kann mit Hilfe der MI aus MSAs bestimmt werden. Die empirische MI baut dabei auf Frequenzbestimmung auf, sodass unterschiedliche Effekte, wie Phylogenie, *finite-size* und das zugrunde liegende Alignment selbst großen Einfluss auf die Berechnung haben. Um diese Effekte zu quantifizieren und vor allem den *finite-size* Effekt zu kompensieren, wurden in der Vergangenheit verschiedene Korrekturterme für diese Problematiken entwickelt. In dieser Arbeit wurde ein vorgeschlagenes Nullmodell von Hamacher et al. [62] für die ribosomalen Proteine des bakteriellen Ribosoms implementiert und untersucht. Dabei konnte gezeigt werden, dass der *finite-size* Effekt bis zu einer Sequenzanzahl von ca. 200 - 300 Sequenzen eine dominierende Rolle spielt. Das in Kapitel 2 beschriebene Nullmodell hat dabei den Vorteil, dass Effekte wie etwa der Selektionsdruck an jeder Position erhalten bleibt.

Mit Hilfe der MI konnte dann an den Sequenzen der ribosomalen Proteine gezeigt werden, dass koevolutionäre Muster in diesem makromolekularen Komplex nicht naiv zu verstehen sind. Es wurde gezeigt werden, dass sich die meisten koevolutionären Signale durch langreichweitige Interaktionen erklären lassen, hingegen spielen direkte Kontakte zwischen den Residuen nur eine untergeordnete Rolle. Um diese Ergebnisse weiter zu interpretieren und eventuell sogar Liganden induzierte Signale zu extrahieren wurden die MI-Ergebnisse in Netzwerke umgewandelt und wurden anhand topologischer Maße differenziell analysiert. Als weiterer Ansatz wurde die biophysikalische Annotation der Moleküle durch GNM genutzt, um die MI-Analyse molekular zu erklären. Bei der Korrelation der biophysikalischen Annotation mittels GNMs und der MI-Ergebnisse zeigte sich im Ribosom, dass ein starkes Koevolutions-Signal keine ähnlich starke Entsprechung in der FN von GNMs hat. Die thermodynamische Analyse mit Hilfe der SCPCP zeigte für die große ribosomale Untereinheit abweichende thermodynamische Einflüsse bei der Assemblierung dieses makromolekularen Komplexes von vorläufigen experimentellen Studien. Die Unterschiede zwischen unseren Vorhersagen für *T. thermophilus* und den oben genannten experimentellen Ergebnissen für *E. coli* bedürfen daher weiterer experimentelle Arbeiten.

Weiterhin ist die mit der MI gefundene Koevolution immer nur auf Grundlage des untersuchten Alphabets zu interpretieren. Daher wurde im Rahmen dieser Arbeit und am Beispiel der HIV-1 Protease untersucht wie sich die Koevolution auf biochemischer Ebene darstellt. Dafür wurden verschiedene Alignments der HIV-1 Protease biochemisch kodiert und dann mit Hilfe der MI analysiert. Dabei wurde deutlich, dass biochemische Koevolution nicht nur durch direkte Interaktion zwischen den koevolvierenden Residuen zustande kommen kann, sondern auch hier durch langreichweitige Interaktionen entsteht. Des Weiteren wurden auch auf biochemischer Ebene schon beobachteten Cluster (das Phylogenetische-Varianz-Cluster und das Drug-Resistanze-Cluster) identifiziert. Da die Alignments der HIV-1 Protease aus einer Datenbank von Patienten stammt, die mit Proteaseinhibitor behandelt wurden, konnten die Alignments nach behandelt und nicht behandelt getrennt werden und so der Effekt von Resistenzmutatio-

nen auf der Ebene biochemischer Koevolution analysiert werden. Dabei konnte ein Unterschied zwischen der Entropie-Verteilung aus AS und biochemischen Eigenschaften identifiziert werden. Dabei wird deutlich, dass viele Resistenz-Mutationen in der Sequenz der HIV-1 Protease zwar zu einer Entropie-Steigerung in der AS-Sequenz führen, die Entropie der biochemisch kodierten Alignments jedoch sehr konserviert vorliegt.

Als letzter Schritt zur Untersuchung molekularer (Ko)evolution wurde in einer interdisziplinäre Zusammenarbeit mit dem Institut für Graphisch-Interaktive Systeme (GRIS) der TU-Darmstadt eine Software entwickelt, die zur Visualisierung und benutzergesteuerten Analyse von mehreren phylogenetischen Bäumen dient.

8 Abkürzungsverzeichnis

AIDS	*Acquired Immune Deficiency Syndrome*
ANM	anisotropes Netzwerk Modell
APC	*Average Product Correction*
AS	Aminosäure
bzw.	beziehungsweise
ca.	circa
cheMI	biochemische *mutual information*
DB	Datenbank
DEMI	Delta-Entropie *mutual information*
d.h.	dass heißt
DNA	Desoxyribonukleinsäure
DRC	*Drug Resistance Cluster*
ENM	Elastisches Netzwerk Modell
ESMI	*EnhancedSubset mututal information*
et al.	et alia
etc.	et cetera
FN	Forbeniusnorm
GB DB	*GenBank database*
GNM	Gauß'sches Netzwerkmodell
GRIS	Institut für Graphisch-Interaktive Systeme
HIV	Humanes Immundefizienz-Virus
HMM	Hidden-Markov Modell
MD	Molekulardynamik
MI	*mutual inforamtion*
mRNA	messenger Ribonukleinsäure
MSA	multiples Sequenzalignment
NCBI	*National Center for Biotechnology Information*
NMA	Normalmodenanalyse
ORMI	*Original mutual information*
PCA	*primary component analysis*
PDB	*Protein Data Bank*
PhVC	phylogenetisches Varianzcluster
PI	Proteaseinhibitor
RCW	*Row Column Weighting*
RNA	Ribonukleinsäure
rRNA	ribosomale Ribonukleinsäure
SCPCP	*self consistent pair contact probability approximation*
SUMI	*Subset mutual information*
SVD	*singular value decomposition*
tRNA	transfer Ribonukleinsäure
VA	*visual analytics*
vergl.	vergleiche
z.B.	zum Beispiel

Literatur

[1] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[2] A. Aleksandrov and T. Simonson. Molecular dynamics simulations of the 30S ribosomal subunit reveal a preferred tetracycline binding site. *JACS*, 130:1114–1115, 2008.

[3] L. B. Almeida. Linear and nonlinear ICA based on mutual information. *Method. Signal Process.*, 84:231–245, 2004.

[4] M. Anisimova and O. Gascuel. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology*, 55(4):539–552, 2006.

[5] O. Aruksakunwong, K. Wittayanarakul, P. Sompornpisut, V. Sanghiran, V. Parasuk, and S. Hannongbua. Structural and dynamical properties of different protonated states of mutant HIV-1 protease complexed with the saquinavir inhibitor studied by molecular dynamics simulations. *J. Mol. Graph. Model*, 25(3):324–332, 2006.

[6] P. Ashorn, T. J. McQuade, S. Thaisrivongs, A. G. Tomasselli, W. G. Tarpley, and B. Moss. An inhibitor of the protease blocks maturation of human and simian immunodeficiency viruses and spread of infection. *PNAS*, 87(19):7472–7476, 1990.

[7] W. R. Atchley, K. R. Wollenberg, W. M. Fitch, W. Terhalle, and A. W. Dress. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol. Biol. Evol.*, 17(1):164–178, 2000.

[8] A. R. Atilgan, S. R. Durrell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80:505–515, 2001.

[9] I. Bahar, A. R. Atilgan, and B. Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3):173–181, 1997.

[10] E. T. Baldwin, T. N. Bhat, B. Liu, N. Pattabiraman, and J. W. Erickson. Structural basis of drug resistance for the V82A mutant of HIV-1 proteinase. *Nat. Struct. Biol.*, 2(3):244–249, 1995.

[11] A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, 5(2):101–113, 2004.

[12] A. Bashan and A. Yonath. Correlating ribosome function with high-resolution structures. *Cell Trends in Microbiology*, 16:326–335, 2008.

[13] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. GenBank. *Nucleic Acids Res.*, 33(Database Issue):D34–D38, 2005.

[14] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. GenBank. *Nucleic Acids Res.*, 36(Database issue):D25–D30, 2008.

[15] P. Boba, P. Weil, F. Hoffgaard, and K. Hamacher. Co-evolution in HIV enzymes. *Proceedings of Bioinformatics*, pages 39–47, 2010.

[16] S. Bremm, T. von Landesberger, M. Hess, T. Schreck, P. Weil, and K. Hamacher. Interactive visual comparison of multiple trees. In *IEEE Conference on Visual Analytics Science and Technology (VAST2011)*, 2011.

[17] D. E. Brodersen, W. M. Clemons, A. P. Carter, R. J. Morgan-Warren, B. T. Wimberly, and V. Ramakrishnan. The structural basis for the action of the antibiotics tetracycline, pactamycin, and hygromycin B on the 30S ribosomal subunit. *Cell*, 103(7):1143–1154, 2000.

[18] B. Brooks and M. Karplus. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *PNAS*, 80(21):6571–6575, 1983.

[19] D. Bulkley, C. A. Innis, G. Blaha, and T. A. Steitz. Revisiting the structures of several antibiotics bound to the bacterial ribosome. *PNAS*, 107:17158–17163, 2010.

[20] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.*, 10(3):186–198, 2009.

[21] L. Burger and E. van Nimwegen. Disentangling direct from indirect Co-Evolution of residues in protein alignments. *PLoS Comput. Biol.*, 6(1):e1000633, 2010.

[22] C. M. Buslje, J. Santos, J. M. Delfino, and M. Nielsen. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics*, 25(9):1125–1131, 2009.

[23] A. J. Butte and I. S. Kohane. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, 4:415–426, 2000.

[24] G. J. Caporaso, S. Smit, B. C. Easton, L. Hunter, G. A. Huttley, and R. Knight. Detecting coevolution without phylogenetic trees? tree-ignorant metrics of coevolution perform as well as tree-aware metrics. *BMC Evol. Biol.*, 8:327, 2008.

[25] A. P. Carter, W. M. Clemons, D. E. Brodersen, R. J. Morgan-Warren, B. T. Wimberly, and V. Ramakrishnan. Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature*, 407(6802):340–348, 2000.

[26] J. Castresana. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, 17(4):540–552, 2000.

[27] L. Chen, A. Perlina, and C. J. Lee. Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. *J. Virol.*, 78(7):3722–3732, 2004.

[28] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res.*, 31(13):3497–3500, 2003.

[29] G. A. Churchill. Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.*, 51(1):79–94, 1989.

[30] G. A. Churchill. Hidden Markov chains and the analysis of genome structure. *Computers & Chemistry*, 16(2):107–115, April 1992.

[31] M. Clamp, J. Cuff, S. M. Searle, and G. J. Barton. The Jalview Java alignment editor. *Bioinformatics*, 20(3):426–427, 2004.

[32] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[33] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[34] Q. Cui, R. K.-Z. Tan, S. C. Harvey, and D. A. Case. Low-Resolution molecular dynamics simulations of the 30S ribosomal subunit. *Multiscale Modeling & Simulation*, 5(4):1248–1263, 2006.

[35] S. de Meyer, A. Hill, G. Picchio, R. de Masi, E. de Paepe, and M.-P. de Béthune. Influence of baseline protease inhibitor resistance on the efficacy of darunavir/ritonavir or lopinavir/ritonavir in the TITAN trial. *J. Acquir. Immune. Defic. Syndr.*, 49(5):563–564, 2008.

[36] S. de Meyer, T. Vangeneugden, B. van Baelen, E. de Paepe, H. van Marck, G. Picchio, E. Lefebvre, and M.-P. de Béthune. Resistance profile of darunavir: combined 24-week results from the POWER trials. *AIDS Res. Hum. Retroviruses*, 24(3):379–388, 2008.

[37] A. Dereeper, V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J.-F. Dufayard, S. Guindon, V. Lefort, M. Lescot, J.-M. Claverie, and O. Gascuel. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.*, 36(Web Server):W465–W469, 2008.

[38] D. Descamps, S. Lambert-Niclot, A.-G. Marcelin, G. Peytavin, B. Roquebert, C. Katlama, P. Yeni, M. Felices, V. Calvez, and F. Brun-Vézinet. Mutations associated with virological response to darunavir/ritonavir in HIV-1-infected protease inhibitor-experienced patients. *J. Antimicrob. Chemother.*, 63(3):585–592, 2009.

[39] J. A. Dunkle, L. Xiong, A. S. Mankin, and J. H. D. Cate. Structures of the *escherichia coli* ribosome with antibiotics bound near the peptidyl transferase center explain spectra of drug action. *PNAS*, 107:17152–17157, 2010.

[40] S. D. Dunn, L. M. Wahl, and G. B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, 2008.

[41] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797, 2004.

[42] B. Efron. Computers and the theory of statistics: Thinking the unthinkable. *SIAM Review*, 21(4):460–480, 1979.

[43] J. Ellson, E. Gansner, L. Koutsofios, S. North, G. Woodhull, Short Description, and Lucent Technologies. Graphviz – open source graph drawing tools. In *Lecture Notes in Computer Science*, pages 483–484. Springer-Verlag, 2001.

[44] P. Erdös and A. Rényi. On random Graphs I. *Publicationes Mathematicae*, 6:290–297, 1959.

[45] J. Felsenstein. PHYLIP – phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.

[46] R. D. Finn, J. Clements, and S. R. Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, 39(Web Server issue):W29–W37, 2011.

[47] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H.-R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman. The Pfam protein families database. *Nucleic Acids Re.*, 36(Database):D281–D288, 2007.

[48] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H.-R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman. The Pfam protein families database. *Nucleic Acids Res.*, 36(Database issue):D281–D288, 2008.

[49] G. E. Fox, E. Stackebrandt, R. B. Hespell, J. Gibson, J. Maniloff, T. A. Dyer, R. S. Wolfe, W. E. Balch, R. S. Tanner, L. J. Magrum, L. B. Zablen, R. Blakemore, R. Gupta, L. Bonen, B. J. Lewis, D. A. Stahl, K. R. Luehrsen, K. N. Chen, and C. R. Woese. The phylogeny of prokaryotes. *Science*, 209(4455):457–463, 1980.

[50] L. C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1978.

[51] H. Gao, Y. Dou, J. Yang, and J. Wang. New methods to measure residues coevolution in proteins. *BMC Bioinformatics*, 12:206, 2011.

[52] R. A. Garrett. *The Ribosome: Structure, Function, Antibiotics, and Cellular Interactions*. AMS Press, Washington, D.C., 2000.

[53] G. B. Gloor, L. C. Martin, L. M. Wahl, and S. D. Dunn. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry*, 44(19):7156–7165, 2005.

[54] N. Go, T. Noguti, and T. Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *PNAS*, 80(12):3696–3700, Jun 1983.

[55] G. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14:403–420, 1970.

[56] R. Gouveia-Oliveira and A. G. Pedersen. Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. *Algorithms for Molecular Biology*, 2:1–12, 2007.

[57] B. J. Grant, A. P. C. Rodrigues, K. M. ElSawy, A. J. McCammon, and L. S. D. Caves. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, 22:2695–2696, 2006.

[58] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704, 2003.

[59] T. Haliloglu, I. Bahar, and B. Erman. Gaussian dynamics of folded proteins. *Physical Review Letters*, 79(16):3090–3093, October 1997.

[60] D. B. Hall, J. Baxter, J. Schapiro, C. A. B. Boucher, C. Tilke, and J. Scherer. Mutations 24I, 50L/V, 54L, and 76V, selected by other protease inhibitors, predict durable response to tipranavir in treatment experienced patients when two or more are present. In *XVII International HIV Drug Resistance Workshop*, 2008.

[61] K. Hamacher. Adaptive extremal optimization by detrended fluctuation analysis. *J. Comp. Phys.*, 227(2):1500–1509, 2007.

[62] K. Hamacher. Relating sequence evolution of HIV1-protease to its underlying molecular mechanics. *Gene*, 422:30–36, 2008.

[63] K. Hamacher. Temperature dependence of fluctuations in HIV1-protease. *Eur. Biophys. J.*, 2009.

[64] K. Hamacher. Efficient perturbation analysis of elastic network models - Application to acetylcholinesterase of *t. californica*. *Journal of Computational Physics*, 229(19):7309–7316, 2010.

[65] K. Hamacher and J. Trylska J. A. McCammon. Dependency Map of Proteins in the Small Ribosomal Subunit. *PLoS*, 2(2):e10, 2006.

[66] M. Herold and K. H. Nierhaus. Incorporation of six additional proteins to complete the assembly map of the 50S subunit from *escherichia coli* ribosomes. *The Journal of Biological Chemistry*, 262(18):8826–8833, June 1987.

[67] K. E. Hild, D. Erdogmus, and J. Principe. Blind source separation using Renyi's mutual information. *Signal Process. Lett.*, 8:174–176, 2001.

[68] F. Hoffgaard, P. Weil, and K. Hamacher. BioPhysConnectoR: Connecting sequence information and biophysical models. *BMC Bioinformatics*, 11:199, 2010.

[69] S. Holmes. Bootstrapping phylogenetic trees: Theory and methods. *Stat. Sci.*, 18(2):241–255, 2003.

[70] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, B. A. Cuche, E. de Castro, C. Lachaize, P. S. Langendijk-Genevaux, and C. J. A. Sigrist. The 20 years of PROSITE. *Nucleic Acids Res.*, 36(Database issue):D245–D249, 2008.

[71] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[72] R. Ishima, D. I. Freedberg, Y.-X. Wang, J. M. Louis, and D. A. Torchia. Flap opening and dimer-interface flexibility in the free and inhibitor-bound HIV protease, and their implications for function. *Structure*, 7:1047–1055, 1999.

[73] S. Johnson, J. J. Torres, J. Marro, and M. A. Munoz. The entropic origin of disassortativity in complex networks. In *Phys. Rev. Lett.* [101], page 208701.

[74] V. A. Johnson, F. Brun-Vézinet, B. Clotet, H. F. Günthard, D. R. Kuritzkes, D. Pillay, J. M. Schapiro, and D. D. Richman. Update of the drug resistance mutations in HIV-1: December 2010. *Top HIV Med.*, 18(5):156–163, 2010.

[75] J. M. Keith. *Bioinformatics: Volume I: Data, Sequence Analysis and Evolution*. Springer-Verlag New York, LLC, 2008.

[76] B. Korber, J. Theiler, and S. Wolinsky. Limitations of a molecular clock applied to considerations of the origin of HIV-1. *Science*, 280(5371):1868–1871, 1998.

[77] O. Kurkcuoglu, Z. Kurkcuoglu, P. Doruker, and R. L. Jernigan. Collective Dynamics of the Ribosomal Tunnel Revealed by Elastic Network Modeling. *Proteins*, 75(4):837–845, 2009.

[78] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, 2007.

[79] D. Li, B. Ji, K. Hwang, and Y. Huang. Crucial roles of the subnanosecond local dynamics of the flap tips in the global conformational changes of HIV-1 protease. *J. Phys. Chem. B*, 114(8):3060–3069, 2010.

[80] F. Liu, A. Y. Kovalevsky, Y. Tie, A. K. Ghosh, R. W. Harrison, and I. T. Weber. Effect of flap mutations on structure of HIV-1 protease and inhibition by saquinavir and darunavir. *Journal of Molecular Biology*, 381(1):102–115, August 2008.

[81] Y. Liu, E. Eyal, and I. Bahar. Analysis of correlated mutations in HIV-1 protease using spectral clustering. *Bioinformatics*, 24(10):1243–1250, 2008.

[82] P. Londei. Archaeal ribosomes. In *eLS*. John Wiley & Sons, Ltd, December 2010.

[83] H. Lu, L. Lu, and J. Skolnick. Development of unified statistical potentials describing protein-protein interactions. *Biophys. J.*, 84(3):1895–1901, 2003.

[84] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006):308–312, 2004.

[85] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[86] B. Mahalingam, J. M. Louis, C. C. Reed, J. M. Adomat, J. Krouse, Y. F. Wang, R. W. Harrison, and I. T. Weber. Structural and kinetic analysis of drug resistant mutants of HIV-1 protease. *European Journal of Biochemistry / FEBS*, 263(1):238–245, July 1999.

[87] S. Mahony, P. E. Auron, and P. V. Benos. Inferring protein-DNA dependencies using motif alignments and mutual information. *Bioinformatics*, 23(13):i297–i304, 2007.

[88] C. M. Manaia, B. Hoste, G. M. Carmen, M. Gillis, A. Ventosa, K. Kersters, and M. S. Da Costa. Halotolerant thermus strains from marine and terrestrial hot springs belong to *thermus thermophilus* (ex oshima and imahori, 1974) nom. rev. emend. *Systematic and Applied Microbiology*, 17(4):526–532, 1995.

[89] L. M. Mansky and H. M. Temin. Lower *in vivo* mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J. Virol.*, 69(8):5087–5094, 1995.

[90] A.-G. Marcelin, B. Masquelier, D. Descamps, J. Izopet, C. Charpentier, C. Alloui, M. Bouvier-Alias, A. Signori-Schmuck, B. Montes, M.-L. Chaix, C. Amiel, G. Dos Santos, A. Ruffault, F. Barin, G. Peytavin, M. Lavignon, P. Flandre, and V. Calvez. Tipranavir-ritonavir genotypic resistance score in protease inhibitor-experienced patients. *Antimicrob. Agents. Chemother.*, 52(9):3237–3243, 2008.

[91] L. C. Martin, G. B. Gloor, S. D. Dunn, and L. M. Wahl. Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, 21(22):4116–4124, 2005.

[92] J. L. Marx. New disease baffles medical community. *Science*, 217(4560):618–621, 1982.

[93] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.

[94] T. J. McQuade, A. G. Tomasselli, L. Liu, V. Karacostas, B. Moss, T. K. Sawyer, R. L. Heinrikson, and W. G. Tarpley. A synthetic HIV-1 protease inhibitor with antiviral activity arrests HIV-like particle maturation. *Science*, 247(4941):454–456, 1990.

[95] C. Micheletti, J. R. Banavar, and A. Maritan. Conformations of proteins in equilibrium. *Phys. Rev. Lett.*, 87(8):1–4, 2001.

[96] E. H. Moore. *Bull. Am. Math. Soc.*, 26:394–395, 1920.

[97] V. I. Morariu, B. V. Srinivasan, V. C. Raykar, R. Duraiswami, and L. S. Davis. Automatic online tuning for fast gaussian summation. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[98] T. Munzner, F. Guimbretière, S. Tasiran, L. Zhang, and Y. Zhou. TreeJuxtaposer: scalable tree comparison using focus+context with guaranteed visibility. *ACM Trans. Graph.*, 22:453–462, 2003.

[99] F. Murtagh. Multidimensional clustering algorithms. *Compstat. Lectures*, 4, 1985.

[100] L. K. Naeger and K. A. Struble. Food and drug administration analysis of tipranavir clinical resistance in HIV-1-infected treatment-experienced patients. *AIDS*, 21(2):179–185, 2007.

[101] M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89(20):208701, 2002.

[102] L. K. Nicholson, T. Yamazaki, D. A. Torchia, S. Grzesiek, A. Bax, S. J. Stahl, J. D. Kaufman, P. T. Wingfield, P. Y.S Lam, P. K. Jadhav, C. N. Hodge, P. J. Domaille, and C. Chang. Flexibility and function in HIV-1 protease. *Nature Structural Biology*, 2:274–280, 1995.

[103] M. Osadchy and R. Kolodny. Maps of protein structure space reveal a fundamental relationship between protein structure and function. *PNAS*, 108(30):12301–12306, 2011.

[104] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[105] C. Pan, J. Kim, L. Chen, Q. Wang, and C. Lee. The HIV positive selection mutation database. *Nucleic Acids Res.*, 35(Database issue):D371–D375, 2007.

[106] I. Pellegrin, L. Wittkop, L. M. Joubert, D. Neau, D. Bollens, M. Bonarek, P.-M. Girard, H. Fleury, B. Winters, M.-C. Saux, J.-L. Pellegrin, R. Thiébaut, D. Breilh, and A. N. R. S. Aquitaine Cohort. Virological response to darunavir/ritonavir-based regimens in antiretroviral-experienced patients (PREDIZISTA study). *Antivir. Ther.*, 13(2):271–279, 2008.

[107] R. Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(03):406–413, 1955.

[108] D. T. Pham. Blind separation of instantaneous mixture of sources via the Gaussian mutual information criterion. *Signal Process.*, 81:855–870, 2001.

[109] M. Pioletti, F. Schlünzen, J. Harms, R. Zarivach, M. Glühmann, H. Avila, A. Bashan, H. Bartels, T. Auerbach, C. Jacobi, T. Hartsch, A. Yonath, and F. Franceschi. Crystal structures of complexes of the small ribosomal subunit with tetracycline, edeine and IF3. *The EMBO Journal*, 20(8):1829–1839, 2001.

[110] D. D. Pollock, W. R. Taylor, and N. Goldman. Coevolving protein residues: Maximum likelihood identification and relationship to structure. *J. Mol. Biol.*, 287:187–198, 1999.

[111] P. Puigbò, S. Garcia-Vallvé, and J. O. McInerney. TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinformatics*, 23(12):1556–1558, 2007.

[112] V. Pulim, J. Bienkowska, and B. Berger. LTHREADER: prediction of extracellular ligand-receptor interactions in cytokines using localized threading. *Protein Sci.*, 17(2):279–292, 2008.

[113] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.

[114] V. Ramakrishnan. Ribosome structure and the mechanism of translation. *Cell*, 108(4):557–572, 2002.

[115] S.-Y. Rhee, M. J. Gonzales, R. Kantor, B. J. Betts, J. Ravela, and R. W. Shafer. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.*, 31(1):298–303, 2003.

[116] S.-Y. Rhee, R. Kantor, D. A. Katzenstein, R. Camacho, L. Morris, S. Sirivichayakul, L. Jorgensen, L. F. Brigido, J. M. Schapiro, R. W. Shafer, and International Non Subtype B HIV-1 Working Group. HIV-1 pol mutation frequency by subtype and treatment experience: extension of the HIVseq program to seven non-b subtypes. *AIDS*, 20(5):643–651, 2006.

[117] E. Roberts, A. Sethi, J. Montoya, C. R. Woese, and Z. Luthey-Schulten. Molecular signatures of ribosomal evolution. *PNAS*, 105(37):13953–13958, 2008.

[118] J. Romanowska, J. A. McCammon, and J. Trylska. Understanding the origins of bacterial resistance to aminoglycosides through molecular dynamics mutational study of the ribosomal A-site. *PLoS Comput. Biol.*, 7(7):e1002099, 2011.

[119] F. Schlünzen, D. N. Wilson, P. Tian, J. M. Harms, S. J. McInnes, H. A. S. Hansen, R. Albrecht, J. Buerger, S. M. Wilbanks, and P. Fucini. The binding mode of the trigger factor on the ribosome: implications for protein folding and SRP interaction. *Structure*, 13(11):1685–1694, 2005.

[120] A. Senes, M. Gerstein, and D. M. Engelman. Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J. Mol. Biol.*, 296(3):921–936, Feb 2000.

[121] R. W. Shafer. Rationale and uses of a public HIV drug-resistance database. *J. Infect. Dis.*, 194 Suppl 1:S51–S58, 2006.

[122] R. W. Shafer, D. R. Jung, and B. J. Betts. Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries. *Nat. Med.*, 6(11):1290–1292, 2000.

[123] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423, 1948.

[124] P. S. Soltis and D. E. Soltis. Applying the bootstrap in phylogeny reconstruction. *Statist. Sci.*, 18(2):256–267, 2003.

[125] E. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman, and R. Durbin. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.*, 26(1):320–322, 1998.

[126] F. Sorrentino, M. di Bernardo, G. Huerta Cuellar, and S. Boccaletti. Synchronization in weighted scale-free networks with degree–degree correlation. *Physica D*, pages 123–129, 2006.

[127] M. A. Steel and D. Penny. Distributions of tree comparison metrics–some new results. *Systematic Biology*, 42(2):126–141, 1993.

[128] S. Tayefeh, T. Kloss, G. Thiel, B. Hertel, A. Moroni, and S. M. Kast. Molecular dynamics simulation of the cytosolic mouth in Kcv-type potassium channels. *Biochemistry*, 46(16):4826–4839, 2007.

[129] W. R. Taylor. The classification of amino acid conservation. *Journal of Theoretical Biology*, 119(2):205–218, 1986.

[130] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22(22):4673–4680, 1994.

[131] M. M. Tirion. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.*, 77(9):1905, 1996.

[132] G. E. Tusnády and I. Simon. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol*, 283(2):489–506, Oct 1998.

[133] D. Voet and J. G. Voet. *Biochemie*. VCH, 1994.

[134] A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton. Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191, 2009.

[135] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *PNAS*, 106(1):67–72, 2009.

[136] P. Weil, F. Hoffgaard, and K. Hamacher. Estimating sufficient statistics in co-evolutionary analysis by mutual information. *Computational Biology and Chemistry*, 33(6):440–444, 2009.

[137] S. Weissgraeber, F. Hoffgaard, and K. Hamacher. Structure-based, biophysical annotation of molecular coevolution of acetylcholinesterase. *Proteins*, 79(11):3144–3154, 2011.

[138] D. N. Wilson, J. M. Harms, K. H. Nierhaus, F. Schlünzen, and P. Fucini. Species-specific antibiotic-ribosome interactions: implications for drug development. *Biological Chemistry*, 386(12):1239–1252, 2005.

[139] A. Wlodawer and J. W. Erickson. Structure-based inhibitors of HIV-1 protease. *Annu. Rev. Biochem.*, 62:543–585, 1993.

[140] A. Wlodawer and J. Vondrasek. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Annu. Rev. Biophys. Biomol. Struct.*, 27:249–284, 1998.

[141] C. R. Woese, L. J. Magrun, R. Gupta, R. B. Siegel, D. A. Stahl, J. Kop, N. Crawford, J. Brosius, R. Gutell, J. J. Hogan, and H. F. Noller. Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res.*, 8:2275–2283, 1980.

[142] K. R. Wollenberg and W. R. Atchley. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *PNAS*, 97(7):3288–3291, 2000.

[143] D. Wu, P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B. J. Tindall, S. D. Hooper, A. Pati, A. Lykidis, S. Spring, I. J. Anderson, P. D'haeseler, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J.-F. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E. M. Rubin, N. C. Kyrpides, H.-P. Klenk, and J. A. Eisen. A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature*, 462:1056–1060, 2009.

Lebenslauf

Philipp Weil

Geburtsdatum:	13. 07. 1979
Geburtsort:	Berlin
Staatsangehörigkeit:	Deutsch

04/2008 – 07/2012 Promotion an der TU Darmstadt im Fachbereich Biologie
in der AG Computational Biology & Simulation:
„Koevolution in molekularen Komplexen."

10/2001 – 11/2007 Studium der Biologie an der TU Darmstadt (Diplom)
Diplomarbeit in der Genetik:
„Die Detektion des trypanosomalen Oberflächenproteins VSG durch einen Aptamer-basierten Biosensor."

10/2000 – 08/2001 Zivildienst in der Jugendherberge Haus Heliand

06/1990 – 06/2000 Besuch des Gymnasiums Oberursel (Abitur)

Publikationen

- P. Weil, F. Hoffgaard, K. Hamacher. *Estimating Sufficient Statistics in Co-Evolutionary Analysis by Mutual Information*, Computational Biology and Chemistry 33(6):440-444, 2009.
- P. Boba, P. Weil, F. Hoffgaard, K. Hamacher. *Co-Evolution in HIV Enzymes*, Proc. of Bioinformatics 2010, p. 39-47, A. Fred, J. Filipe, H. Gamboa (eds.)
- F. Hoffgaard, P. Weil, K. Hamacher. *BioPhysConnectoR: Connecting Sequence Information and Biophysical Models*, BMC Bioinformatics, 11:199, 2010.
- S. Bremm, T. von Landesberger, M. Hess, T. Schreck, P. Weil, K. Hamacher, *Interactive Visual Comparison of Multiple Trees*, IEEE Conference on Visual Analytics Science and Technology (VAST2011), accepted

DANKE.[9]

[9] Im Speziellen an Patrik Boba, Sebastian Bremm, Miriam Carbon-Mangels, Henrik Cordes, Kay Hamacher, Martin Hess, Franziska Hoffgaard, Sven Jager, Frank Keul, Sabine Knorr, Tatiana von Landesberger, Tobias Schreck, Stephanie Weil, Anne Weil, Peter Weil, Max Weil, Stephanie Weißgräber. Und natürlich der AG-WG und allen Anderen, die hier nicht im einzelnen genannt wurden.