

Institut für Visualisierung und Interaktive Systeme

Universität Stuttgart
Universitätsstraße 38
D-70569 Stuttgart

Diplomarbeit

Visualisierung der Unsicherheit der Sekundärstruktur von Proteinen

Matthias Braun

Studiengang:	Informatik
Prüfer/in:	Prof. Dr. Thomas Ertl
Betreuer/in:	Dr. Michael Krone, Karsten Schatz, M.Sc. Dipl.-Inf. Christoph Schulz
Beginn am:	29. September 2016
Beendet am:	31. März 2017
CR-Nummer:	G.3, I.3.3, I.3.7, J.3

Kurzfassung

In allen Information und Daten ist Unsicherheit enthalten. Visualisierungen von Informationen oder Daten entsprechen der Realität deshalb genauer, wenn sie um eine Repräsentation der Unsicherheit ergänzt werden. In dieser Diplomarbeit sind Sekundärstrukturen, welche bei Proteinen durch die Faltung der Aminosäurenkette entstehen, die Datengrundlage. Die Beschreibung der korrekten Sekundärstruktur ist für das Verständnis der Funktion eines Proteins von grundlegender Bedeutung. Unterschiedliche Verfahren weichen allerdings in ihren Zuordnungen von Sekundärstrukturen teilweise deutlich voneinander ab. Ausgehend davon ergab sich als Aufgabe die Visualisierung der Unsicherheit der Sekundärstruktur von Proteinen. Dafür wurde ein Modell der Unsicherheit für abweichende Sekundärstruktur-Zuordnungen entwickelt. Auf der Ebene der Verfahren wurde die Zuordnungsunsicherheit durch ein Schwellenwert-Kriterium beschrieben. Anhand von Diskrepanz-Matrizen wurden die Zuordnungen der einzelnen Verfahren verglichen. Für jede Aminosäure wurden infolgedessen Zuordnungswahrscheinlichkeiten berechnet. Diese wurden wiederum auf einen Unsicherheitswert je Aminosäure reduziert. Die Zuordnungswahrscheinlichkeiten werden im Sequenz-Diagramm mittels Struktur-Morphing und Farbinterpolation im HSL-Farbraum dargestellt, die Unsicherheits- und Schwellenwerte über Säulen-Diagramme. In der Cartoon-Darstellung kann der Unsicherheitswert in Form von Geometrieverzerrung oder Konturen dargestellt werden. Die Repräsentation der Strukturtyp-Wahrscheinlichkeiten erfolgt über Screen-Door-Transparency. Die Betrachtung beider Ansichten nebeneinander ermöglicht eine Ergänzung der unterschiedlichen Unsicherheitsdarstellungen. Darüber hinaus wird eine vollständigere und genauere Wiedergabe der vorhandenen Daten bezüglich der Sekundärstruktur-Zuordnung erreicht.

Inhaltsverzeichnis

1	Einleitung	1
2	Grundlagen	5
2.1	Strukturbiologische Grundlagen	5
2.1.1	Verfahren zur Bestimmung der Sekundärstruktur	6
2.2	Konzept der Unsicherheit	8
3	Verwandte Arbeiten	11
3.1	Vergleich von Strukturzuordnungen	11
3.2	Struktur- und Sequenz-Visualisierung	12
3.3	Visualisierung von Unsicherheit	14
4	Modell für Unsicherheit	19
4.1	Genauigkeit der Atom-Koordinaten	20
4.2	Datenerfassung	21
4.2.1	Dateiformate	22
4.2.2	Ausgabe zusätzlicher Werte bei STRIDE	25
4.2.3	Einlesen der Strukturzuordnungen	26
4.2.4	Ausgabeformat	27
4.3	Strukturwahrscheinlichkeiten auf Ebene der Verfahren	28
4.3.1	Initialisierung der Wahrscheinlichkeiten	30
4.3.2	Unsicherheitskriterium für STRIDE	30
4.4	Diskrepanzkriterien auf Ebene der Aminosäuren	37
4.4.1	Generelle Strukturunterschiede	37
4.4.2	STRIDE - DSSP	40
4.4.3	PROMOTIF	42
4.4.4	Manuelle Zuordnung	43
4.4.5	Berechnung der Strukturtyp-Wahrscheinlichkeiten	44
4.5	Reduktion auf einen Unsicherheitswert	45

5	Visualisierung	47
5.1	Analyse geeigneter Darstellungskonzepte	47
5.2	Sequenz-Diagramm	51
5.2.1	Struktur-Morphing	52
5.2.2	Farbinterpolation im Balkendiagramm	53
5.2.3	Säulendiagramm für Unsicherheits- und Schwellenwerte	57
5.3	Cartoon-Darstellung	57
5.3.1	Geometrieverzerrung	58
5.3.2	Konturen	61
5.3.3	Screen-Door-Transparency	62
6	Ergebnisse	65
6.1	Visualisierungen der Unsicherheit	65
6.1.1	Sequenz-Diagramm	66
6.1.2	Cartoon-Darstellung	67
6.1.3	Aufgeteilte Ansicht	70
6.2	Bewertung und Ausblick	71
6.2.1	Performanz	71
6.2.2	Unsicherheitsmodell	73
6.2.3	Evaluierung der Visualisierungen	74
7	Zusammenfassung	79
	Abbildungsverzeichnis	81
	Tabellenverzeichnis	82
	Literaturverzeichnis	83

Aus Gründen der besseren Lesbarkeit wird auf die gleichzeitige Verwendung männlicher und weiblicher Sprachformen verzichtet. Sämtliche Personenbezeichnungen gelten gleichwohl für beiderlei Geschlecht.

Kapitel 1

Einleitung

In den letzten Jahren wurden sowohl die Quantifizierung als auch die Visualisierung von Unsicherheit zu einem immer bedeutenderen Forschungsgebiet.¹ Demnach ist es nicht nur wichtig, die in jeglicher Information enthaltene Unsicherheit zu quantifizieren, sondern auch die Position und das Ausmaß der Unsicherheit durch eine geeignete Visualisierung zu vermitteln. Die Visualisierung von Daten entspricht der Realität genauer, wenn sie um eine Repräsentation der Unsicherheit ergänzt wird. Die genauere Abbildung der Realität hilft dabei zu verstehen, welche Teile der Informationen oder Daten genau, vollständig, konsistent oder sicher sind und welche nicht. Damit bildet die Visualisierung der Unsicherheit die Grundlage für gut fundierte Entscheidungsprozesse und stärkt das Vertrauen in das bei der Analyse der Daten gewonnene Resultat. Inzwischen existieren für die unterschiedlichsten Anwendungsfälle spezifische und individuelle Modelle, um Unsicherheit zu quantifizieren. Es gibt auch zahlreiche Methoden, um Unsicherheit zu visualisieren.

Das Thema dieser Diplomarbeit ist die Visualisierung der Unsicherheit der Sekundärstruktur von Proteinen. Ausgangspunkt sind hier also Proteine, welche in den Zellen aller Organismen vorkommen und dort eine Vielzahl an Aufgaben erfüllen. Diese reichen vom Katalysieren chemischer Reaktionen bis hin zum Transport anderer Moleküle. Aufgebaut sind Proteine aus einer Aneinanderreihung von Aminosäuren. Die Sekundärstruktur entsteht, wenn sich die Aminosäurekette, die Primärstruktur eines Proteins, faltet. Das Protein bildet seine funktionale Struktur durch die räumliche Faltung aus. Die Beschreibung der korrekten dreidimensionalen Anordnung der Sekundärstruktur ist deshalb für das Verständnis der Funktion eines Proteins von grundlegender Bedeutung. Bei den unterschiedlichen Sekundärstrukturtypen (z.B. α -Helices und β -Faltblätter), welche verschiedene lokale Strukturen repräsentieren, handelt es sich um ein Klassifikationsschema. Die Sekundärstrukturtypen sind über Kriterien definiert, sowohl auf der Grundlage von Wasserstoff-Brückenbindungen als auch anhand der

¹Beispielsweise: *SFB-TRR 161* - Teilprojekt A01: Quantifizierung und Analyse der Unsicherheit im Bereich Visual Computing. Webseite: <http://www.trr161.de> oder *SFB 716* - Teilprojekt D.4: Interaktive Visualisierung dynamischer, komplexer Eigenschaften von Protein-Lösungsmittel-Systemen. Webseite: <http://www.sfb716.uni-stuttgart.de>

dreidimensionalen Positionen der Atome. Für die Bestimmung der Sekundärstruktur eines Proteins stehen neben der Option der manuellen Zuordnung durch einen Experten auch eine Vielzahl automatisierter Verfahren zur Verfügung. Die als Referenz bei der Zuordnung von Sekundärstrukturen angesehene manuelle Zuordnung ist allerdings sehr aufwendig und wird somit nur bei den wenigsten Proteinen angewendet. Deswegen werden dafür meist Programme verwendet, wie beispielsweise STRIDE oder DSSP. Die resultierenden Zuordnungen dieser Programme weichen jedoch bis zu einem gewissen Grad voneinander ab. Außerdem unterscheiden sie sich teilweise deutlich von der manuell bestimmten Sekundärstruktur. Unterschiedliche Zuordnungen von Sekundärstrukturen erzeugen außerdem theoretische Problemstellungen und praktische Einschränkungen bei der Analyse und Verwendung von biologischen Makromolekülen wie Proteinen. Probleme können so beispielsweise bei Benutzerstudien, bei Vorhersagen durch *Ab-initio*-Methoden oder bei spezifischen biophysiochemikalischen Charakterisierungen auftreten [Roc14]. Die Abweichungen in der Strukturzuordnung lassen zudem Zweifel über die Qualität der verwendeten Zuordnungsverfahren aufkommen und auch darüber, welche der Zuordnungsergebnisse verwendet werden sollen. Ausgehend davon ergibt sich die nachfolgend formulierte Aufgabenstellung dieser Diplomarbeit.

Aufgabenstellung

Das Ziel dieser Diplomarbeit ist es, die Ergebnisse verschiedener Programme oder Algorithmen zur Extraktion der Sekundärstruktur (z.B. STRIDE [FA95] und DSSP [KS83]) miteinander und mit der manuell bestimmten Sekundärstruktur zu vergleichen. Aus den Ergebnissen soll, basierend auf der Übereinstimmung der Resultate, ein Unsicherheitsfaktor pro Aminosäure berechnet werden. Diese Unsicherheitsfaktoren sollen sowohl in der dreidimensionalen Cartoon-Darstellung als auch im zweidimensionalen Sequenz-Diagramm geeignet visualisiert werden. Es ist darauf zu achten, dass die Analyse der Daten durch die Darstellung der Unsicherheit nicht eingeschränkt wird. Beispielsweise ist die Verwendung von Farbe zur Darstellung der Unsicherheit nur bedingt geeignet, da Farbe in der Molekülvisualisierung üblicherweise zur Darstellung physikochemischer Eigenschaften verwendet wird.

Im weiteren Verlauf der Arbeit soll die Unsicherheitsvisualisierung geeignet erweitert und angewendet werden. Möglich ist zum Beispiel der Einsatz zur Qualitätsanalyse anderer Algorithmen zur Sekundärstrukturbestimmung, beispielsweise des PROSIGN-Algorithmus [HSP+08]. Denkbar wäre auch, die Unsicherheit einzelner Algorithmen, welche meist feste Schwellenwerte zur Klassifikation verwenden, genauer darzustellen. Es kann z.B. angenommen werden, dass Werte in der Nähe eines Schwellenwertes weniger sicher sind. Die Unsicherheitsvisualisierung kann auch auf dynamische Daten angewendet werden. Hier kann dargestellt werden, wie sicher das Vorliegen eines bestimmten Sekundärstrukturelements ist, indem seine Stabilität über die Zeit hinweg angegeben wird.

Die Implementierung wird in das von Grottel et al. [GKM+15] entwickelte MegaMol™-Framework integriert und erfolgt in der Programmiersprache C/C++. Dieses Visualisierungs-Framework beinhaltet verschiedene Verfahren zur Visualisierung und Analyse von Proteinen.

Ein Beispiel ist die GPU-beschleunigte Visualisierung der Sekundärstruktur von Proteinen. Als Grafik-Programmierschnittstelle wird die plattformunabhängige OpenGL (Open Graphics Library) in Verbindung mit GLSL (OpenGL Shading Language) verwendet.²

²Webseite von OpenGL mit umfangreicher Dokumentation: <https://www.opengl.org/>

Kapitel 2

Grundlagen

Zu Beginn der Arbeit wird in diesem Kapitel auf die für das weitere Verständnis notwendigen Grundlagen der Strukturbiologie von Proteinen eingegangen und das Konzept der Unsicherheit beschrieben. Vorausgesetzt werden grundlegende Kenntnisse in der Wahrscheinlichkeitsrechnung, der linearen Algebra und der Grafikprogrammierung.

2.1 Strukturbiologische Grundlagen

Proteine sind an den meisten biologischen Prozessen beteiligt und deshalb für alles Leben auf der Erde von enormer Bedeutung. Proteine sind Makromoleküle, sogenannte Polypeptide, welche aus einer oder mehreren Ketten von Aminosäuren bestehen. Der verbleibende Rest einer Aminosäure im Protein, nachdem diese eine Peptidbindung unter Abspaltung von H_2O eingegangen ist, wird üblicherweise als *Residue* bezeichnet. Da es im Deutschen keine entsprechende Bezeichnung gibt, bezieht sich der Begriff *Aminosäure* im Weiteren auch auf in Proteinen gebundene Aminosäuren. Die Funktion eines Proteins hängt von seiner räumlichen Faltung ab, welche somit von besonderer Bedeutung ist. Mit der Sekundärstrukturzuordnung (im Folgenden verkürzt *Zuordnung* genannt) wird die Suche nach strukturell interessanten Stellen innerhalb der dreidimensionalen Anordnung eines Proteins unterstützt. Die Sekundärstruktur dient dazu, die Funktion von Proteinen zu bestimmen, und definiert sich über eine bestimmte lokale räumliche Anordnung der Polypeptidkette. Der Fokus liegt dabei auf der Hauptkette (oder *Backbone*), welche von den kovalent gebundenen Atomen der Aminosäuren gebildet wird. Die in dieser Arbeit behandelten Sekundärstruktur-Zuordnungsverfahren (im Weiteren kurz als *Verfahren* bezeichnet) verwenden eine Teilmenge von acht unterschiedlichen Sekundärstrukturtypen (nachstehend mit *Strukturtypen* abgekürzt). Jeder Aminosäure wird dabei immer genau ein Strukturtyp zugewiesen. Es werden folgende Strukturtypen unterschieden, wobei ggf. in Klammern die dazugehörige englische Bezeichnung mit angegeben ist: π -Helix, α -Helix, 3_{10} -Helix, Umdrehung (Turn), Biegung (Bend), Zufällige Windung oder Krümmung (Coil, Loop), β -Brücke (Bridge), (Erweiterter) β -Strang (Strand). Sind mehrere

β -Stränge nebeneinander angeordnet, bilden sie ein β -Faltblatt (Sheet). Eine detaillierte Beschreibung der einzelnen Strukturtypen erfolgt Abschnitt 4.4.1 im Zusammenhang mit der Betrachtung allgemeiner Strukturunterschiede.

2.1.1 Verfahren zur Bestimmung der Sekundärstruktur

Die im Rahmen dieser Arbeit berücksichtigten Verfahren beruhen bei ihren Zuordnungen neben der Aminosäuresequenz hauptsächlich auf den Koordinaten der Atome. Diese Verfahren extrahieren die Sekundärstrukturen aus den vorhandenen Atom-Koordinaten anhand bestimmter Kriterien. Die Zuordnung von Sekundärstrukturen ist allerdings nicht exakt und eindeutig definiert, wodurch sich die hier zu untersuchenden und zum Teil stark voneinander abweichenden Zuordnungen der verschiedenen Verfahren ergeben. Für die Bestimmung der Atom-Koordinaten wird in den meisten Fällen die nachfolgend beschriebene Röntgenkristallographie verwendet. Von den hier verwendeten Verfahren abzugrenzen sind die sogenannten Vorhersageverfahren. Diese berechnen Sekundärstrukturen ausschließlich auf der Grundlage der Aminosäuresequenz, und somit sind deren Zuordnungen von zusätzlicher Unsicherheit behaftet. Dabei treffen diese Verfahren genau genommen keine Vorhersage. Die Sekundärstrukturen von Proteinen sind keine in der Zukunft liegende Ereignisse, welche vorhergesagt werden können, sondern sie existieren bereits alle, auch wenn sie noch unbekannt sind. Bei den Aussagen zu Sekundärstrukturen handelt sich bei diesen Verfahren also lediglich um Annahmen.

Röntgenkristallographie

Bei der Röntgenkristallographie (oder Röntgenstrahlenstrukturanalyse) wird im ersten Schritt von den zu untersuchenden Protein ein reiner Kristall mit regulärer Struktur hergestellt. Im zweiten Schritt werden mit Röntgenstrahlung einer bestimmten Wellenlänge räumlich variierende Beugungsmuster des Kristallgitters erzeugt und eine dreidimensionale Elektronendichtekarte berechnet. Durch Einpassung der bekannten Aminosäuresequenz des Proteins wird ein Modell der Atom-Positionen erstellt. Davon ausgehend erfolgt in einem iterativen Prozess eine Verfeinerung dieses Modells durch Anpassung der Modell-Parameter. Für eine ausführlichere Betrachtung der durch Röntgenstrahlenstrukturanalyse gewonnenen Molekülstrukturen von Proteinen sei auf Wlodawer et al. [WMDJ08] verwiesen.

PDB

Die *Protein Data Bank* [BWF+00] (im Folgenden kurz: *PDB*) wird seit 1971 in den Brookhaven National Laboratories als Archiv für Strukturdaten biologischer Makromoleküle ständig erweitert. Die PDB wurde im Jahr 2000 frei zugänglich gemacht und ist die zentrale Datenbank für Strukturdaten. Die Datenbank enthält mittlerweile über 128300 Einträge. Jedem

Protein ist dabei eine eindeutige vierstellige PDB-ID zugewiesen. Die in den PDB-Dateien angegebenen umfangreichen Informationen zur Struktur eines Proteins umfassen neben den Atom-Positionen beispielsweise auch die hier relevanten Angaben zur Sekundärstruktur.

Für die automatisierte Berechnung einer Sekundärstruktur wird in den PDB-Dateien das Softwarepaket *PROMOTIF* von Hutchinson und Thornton [HT96] verwendet. Die Berechnung der Sekundärstruktur ist dabei nur ein Teil des Funktionsumfangs von PROMOTIF. Wie in den *Processing Procedures* [Sta14] von worldwidePDB angeführt, wird PROMOTIF dort explizit als bevorzugtes Verfahren genannt.¹ Die Grundlage ist eine modifizierte Implementierung von DSSP (siehe unten). In den PDB-Dateien werden von den umfangreichen und detaillierten Informationen, welche PROMOTIF ausgibt, lediglich diejenigen der unterschiedlichen Helices und der β -Stränge aufgeführt. Die vollständige Ausgabe von PROMOTIF kann man sich beispielsweise bei PDBsum anzeigen lassen.²

Alternativ dazu erfolgt die *manuelle Zuordnung* von Sekundärstrukturen in ca. 1% der PDB-Dateien durch einen Kristallographie-Experten. Sie wird auch als Autoren-Zuordnung bezeichnet, da es sich bei dem Kristallographen meistens um den Autor des jeweiligen PDB-Eintrags handelt. Die so erfolgte visuelle Zuordnung wird bei der Beurteilung von automatisierten Zuordnungsverfahren mangels einheitlich definierter Sekundärstruktur-Definitionen häufig als Referenz verwendet und als „Wahrheit“ angesehen (siehe z.B. STRIDE [FA95]). Diese manuellen Zuordnungen werden in den PDB-Dateien mit einem zusätzlichen Vermerk angegeben.

DSSP

DSSP (Define Secondary Structure of Proteins) wurde von Kabsch und Sander [KS83] vorgestellt und ist eines der ältesten Zuordnungsverfahren. Der Algorithmus basiert auf der Ermittlung von Wasserstoff-Brückenbindungen zwischen Atomen der Hauptkette, welche durch ein elektrostatisches Kriterium definiert sind. Ausgehend davon werden als Grundmuster Umdrehungen und β -Brücken bestimmt. Die anderen Sekundärstrukturen werden dann anhand dieser beiden Grundmuster berechnet und entsprechend einer Aminosäure zugewiesen. Hier wird die neu geschriebene Variante des DSSP-Algorithmus verwendet, welche die bisher vernachlässigten π -Helices besser erkennt [TBB+15].

STRIDE

Dem von Frishman und Argos [FA95] entworfenen Verfahren STRIDE (Structural Identification) liegen für die Bestimmung von Sekundärstrukturen die Fragen zugrunde, welche Kriterien Kristallographen bei der Zuordnung verwenden und wie man diese Kriterien am

¹wwPDB ist die internationale Organisation für die PDB: <http://www.wwpdb.org/>

²Webseite von PDBsum: <https://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=index.html>

besten nachbilden kann. STRIDE verwendet dazu eine ähnliche Definition des Kriteriums für Wasserstoff-Brückenbindungen wie DSSP, beruht allerdings auf einer anderen Funktion zur Berechnung der Bindungsenergien. Zusätzlich fließen bei den Kriterien für α -Helices und β -Brücken noch die aus den Torsionswinkeln ϕ und ψ berechneten Wahrscheinlichkeiten mit ein. Anhand von empirisch bestimmten Schwellenwerten wurden die Zuordnungskriterien an manuellen Zuordnungen optimiert.

2.2 Konzept der Unsicherheit

In der Literatur gibt es sowohl über die Wahrnehmung von Unsicherheit als auch über eine allgemeingültige Repräsentation von Unsicherheit keinen Konsens Pang [Pan01]. Als eine der wenigen definieren Hunter und Goodchild [HG93] Unsicherheit schlüssig als „*das Ausmaß der Unkenntnis über den Umfang von Fehlern, welche dafür verantwortlich ist, dass Ergebnisse und Beobachtungen nicht bedenkenlos akzeptiert werden.*“ Im Allgemeinen wird Unsicherheit als aus vielen Konzepten zusammengesetzt verstanden, welche Genauigkeit, Fehlerfreiheit, Qualität, Fehler, Gültigkeit, Störung, Veränderlichkeit, Vollständigkeit, Vertrauen und Verlässlichkeit umfassen. Bei der hier betrachteten Unsicherheit aufgrund abweichender Zuordnungen von Sekundärstrukturen bei Proteinen handelt es sich, zur Abgrenzung von anderen Arten von Unsicherheit, genauer um *Messunsicherheit* (verkürzt weiterhin als *Unsicherheit* bezeichnet). Der Umgang mit (Mess-)Unsicherheit ist im Standard-Leitfaden für die Berechnung von Unsicherheit, dem „Evaluation of measurement data - Guide to the expression of uncertainty in measurement“ (kurz GUM) [Jcg08] beschreiben. Darin werden vier Möglichkeiten aufgeführt, wie Unsicherheit ausgedrückt werden kann:

- Unter Angabe des geschätzten Mittelwertes und der Standardabweichung kann Unsicherheit *statistisch* beschrieben werden. Die Standardabweichung kann dazu verwendet werden, ein Vertrauensbereich oder eine tatsächliche Verteilung der Daten zu berechnen.
- Über eine Differenz oder einen absoluten *Fehlerwert* innerhalb abgeschätzter Daten oder zwischen einem bekannten korrekten Wert und einer Abschätzung des Wertes.
- Mit der Angabe eines *Bereiches*, welcher ein Intervall beschreibt, innerhalb dessen die Daten liegen müssen. Die Daten können dabei weder statistisch noch über die Fehler-Definition quantifiziert werden.
- Basierend auf *wissenschaftlicher Beurteilung* unter Einbeziehung aller relevanter Informationen sowie von Erfahrung und Wissen, welches über das Verhalten und die Eigenschaften des betrachteten Materials, Instruments oder Prozesses bekannt sind.

In der von Pang, Wittenbrink und Lodha [PWL97] vorgeschlagenen Visualisierungspipeline der Unsicherheit werden die drei wichtigsten Schritte bei der Visualisierung von Unsicherheit

dargestellt, welche letztlich zur Analyse der ausgegebenen Visualisierung führen, siehe Abbildung 2.1. Während der einzelnen Schritte können folgende verschiedene Arten der Unsicherheit entstehen:

- Die bei der Erfassung der Daten entstehende Datenunsicherheit, welche z.B. durch Messfehler, Simulationsfehler, ungeeignete Datenerfassungsmethoden oder durch statistische Schwankungen (bei Beobachtungen oder bei numerischen Modellen) hervorgerufen wird.
- Eine ableitende Unsicherheit, welche sich z.B. durch Resampling, Interpolation, Quantifizierung, Skalierung, Zusammenführung oder Umwandlung der Daten in andere Maßeinheiten ergibt.
- Eine im Zuge der Generierung der Darstellung entstehende Visualisierungsunsicherheit, welche z.B. durch globale Beleuchtung, auf Interreflexion (radiosity) beruhende Abschätzung der Beleuchtung, Approximation bei der Oberflächendarstellung oder Interpolation von Zwischenbildern bei Animationen verursacht wird.

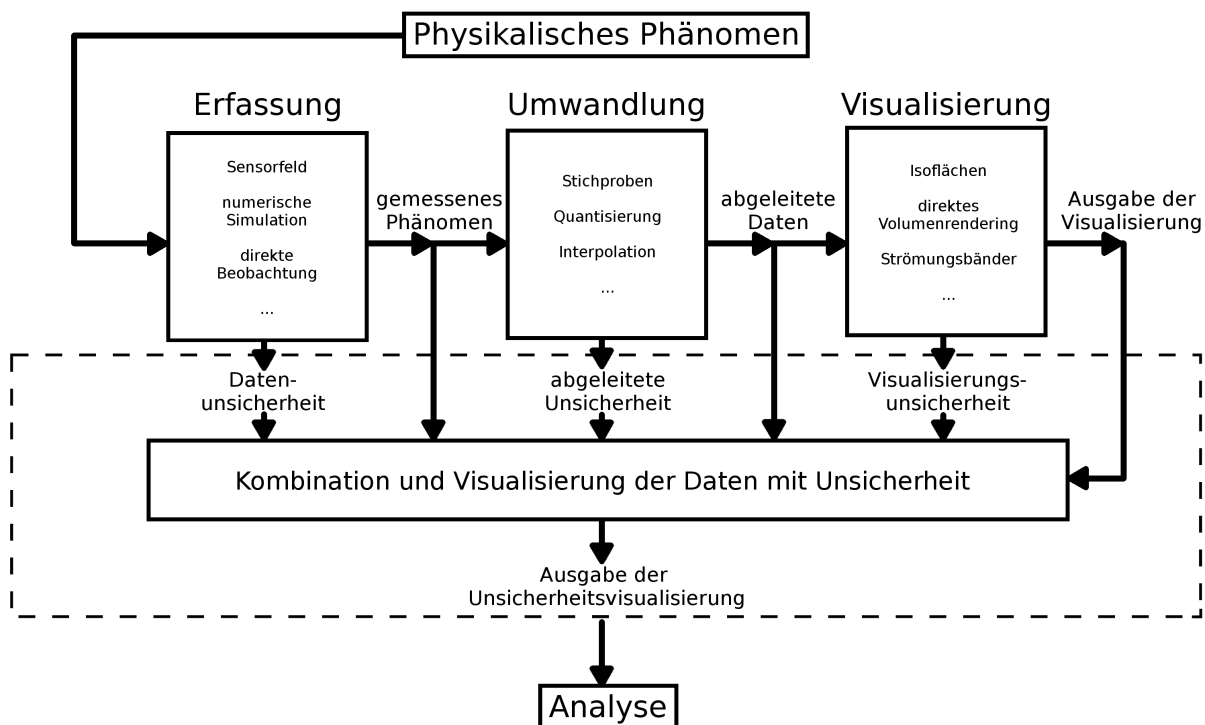


Abbildung 2.1: Visualisierungspipeline der Unsicherheit. Quelle: Abbildung basiert auf [PWL97]

Nach dem Konzept der Fehlerfortpflanzung erfolgt über eine Dimensionsreduktion die Kombination der Unsicherheiten. Die ausgegebene resultierende Unsicherheitsvisualisierung kann dann visuell analysiert werden. Der gestrichelte Bereich in Abbildung 2.1 markiert dabei

2 Grundlagen

diejenigen Themen, deren ausführliche Betrachtung laut der Autoren für die Entwicklung neuer Unsicherheitsvisualisierungen zentral sind. Diese Visualisierungspipeline dient deshalb im Folgenden zur Veranschaulichung der hier angewendeten Vorgehensweise und bildet den Aufbau der vorliegenden Arbeit ab.

Kapitel 3

Verwandte Arbeiten

In diesem Kapitel wird die hier bearbeitete Aufgabe in den aktuellen wissenschaftlichen Stand der vorliegenden Thematik eingeordnet. Da bisher für die Berechnung der Zuordnungsunsicherheit von Sekundärstrukturen kein Modell beschrieben wurde, erfolgt zum einen die Betrachtung von Arbeiten, welche mit verschiedenen Methoden unterschiedliche Zuordnungsverfahren vergleichen. Zum anderen werden Arbeiten sowohl aus dem Bereich der Struktur- und Sequenz-Visualisierung als auch aus dem Bereich der Visualisierung von Unsicherheit erörtert.

3.1 Vergleich von Strukturzuordnungen

Unterschiedliche Zuordnungen von Sekundärstrukturen werden häufig im Zusammenhang mit der Evaluation neuer Zuordnungsverfahren verglichen, z.B. bei STRIDE [FA95], PROSIGN [HSP+08] oder P-SEA [LCPM97]. Zur Evaluierung ihres auf unregelmäßige Strukturen ausgerichteten Verfahrens KAKSI stellen Martin et al. [MLM+05] ihr Verfahren gleich sechs anderen Verfahren gegenüber. Hier dient der globale C_3 -Wert als Vergleichswert, wobei zusätzlich eine Gegenüberstellung anhand des SOV erfolgt.¹ Zur Repräsentation unterschiedlicher Zuordnungen wird die Gegenüberstellung zweier Protein-Strukturen in der Cartoon-Darstellung verwendet. Das Ziel bei Zhang und Sagui [ZS15] war es, die Verfahren STRIDE, DSSP und KAKSI anhand ihrer unterschiedlichen Zuordnungen speziell bei unregelmäßigen Konformationen hin zu vergleichen. Drei Proteine mit unregelmäßiger Struktur dienen dabei als Grundlage. Die statistischen C_3 - und SOV-Werte dienen wieder als Vergleichsmaß. Auch hier werden Zuordnungsunterschiede durch eine Gegenüberstellung in der Cartoon-Darstellung aufgezeigt.

¹Der C_3 -Wert ist ein Vergleichskriterium basierend auf den Sekundärstrukturtypen der einzelnen Aminosäuren. Er gibt die prozentuale Übereinstimmung der identischen Strukturtypen zwischen zwei Sequenzen wieder. Beim Segment *Overlap* handelt es sich um ein auf den Segmenten zweier Sekundärstrukturen basierendes Vergleichsmaß (siehe [RSS94] und [FRZ99]).

Um für Vorhersageverfahren bessere Vergleichszuordnungen von Sekundärstrukturen zu erzeugen, berechnen Colloc'h et al. [CET+93] eine Datenbank mit sogenannten Konsens-Zuordnungen. Dafür werden die Zuordnungen der drei automatisierten Zuordnungsverfahren DSSP, P-Curve und DEFINE für 154 nicht-redundante Proteine kombiniert. Bei einer Reduktion auf drei unterschiedliche Strukturtypen (*Helix*, *Sheet*, *Coil*), ergeben sich somit zehn unterschiedliche Zustände für mögliche Zuordnungen. Einer Aminosäure wird in der resultierenden Konsens-Zuordnung dabei immer derjenige Strukturtyp zugewiesen, welcher am häufigsten an dieser Stelle zugeordnet wurde (schwacher Konsens). Dadurch können Artefakte in den Zuordnungen der einzelnen Verfahren abgeschwächt werden. Für die untersuchten Vorhersageverfahren ergeben die zehn differenzierten Zustände eine bessere Möglichkeit, die angenommene Zuordnung zu überprüfen und zu optimieren. Cuff und Barton [CB99] verwenden die Zuordnungen dreier Verfahren auch zum Zweck einer besseren Evaluation und Optimierung von Vorhersageverfahren. Die untersuchten Vorhersageverfahren erwarten als Eingabe mehrere Sequenzen von Sekundärstruktur-Zuordnungen. Dafür wurden die Verfahren DSSP, STRIDE und DEFINE auch hier statistisch anhand des C_3 -Wertes untersucht. Hier wird zudem festgestellt, dass sich unterschiedliche Vorgehensweisen bei der Reduktion auf die drei üblicherweise verwendeten Strukturtypen *Helix*, *Sheet* und *Coil* auf die erzielte Genauigkeit der Vorhersageverfahren auswirkt. Eine Reduktion der differenzierten Strukturtypen resultiert offensichtlich in zusätzlicher Unsicherheit beim Vergleich von Zuordnungen. Aus diesem Grund wurde hier auf eine solche Reduktion verzichtet und die detaillierte Differenzierung der Strukturtypen beibehalten.

Rocha [Roc14] erforscht die strukturellen Abweichungen der Sekundärstruktur-Zuordnungen unterschiedlicher Verfahren. Beim Vergleich der Verfahren wird dabei kein statistischer Ansatz verwendet. Stattdessen werden konkrete physikochemische Einflüsse wie hydrophobe Wechselwirkungen und nachteilige räumliche Anordnungen auf Aminosäureebene untersucht. Auf der Grundlage seiner Erkenntnisse leitet er drei neue Zuordnungskriterien ab und schlägt ein darauf basierendes Verfahren vor. Hierbei erfolgt die grafische Gegenüberstellung bei der Auswertung von Zuordnungsunterschieden durch parallele Anordnung kurzer eindimensionaler Sequenzabschnitte.

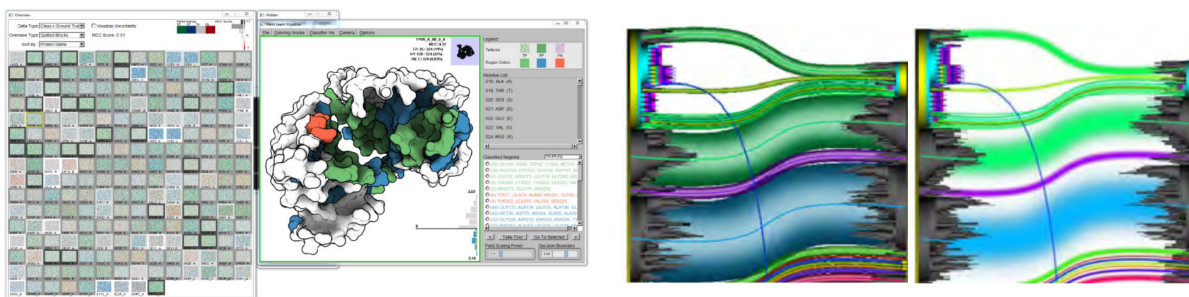
3.2 Struktur- und Sequenz-Visualisierung

Die Arbeit von Kocincová et al. [KJB+17] beschäftigt sich mit der Fragestellung, wie sowohl in drei- als auch in zweidimensionalen Darstellungen Sequenzunterschiede verglichen werden können. Das zugrunde liegende Anwendungsgebiet umfasst hier allerdings nicht verschiedene Zuordnungen von Sekundärstrukturen eines Proteins, sondern geometrische Strukturvergleiche zwischen unterschiedlichen Proteinen. Die entwickelte Darstellung ist in Abbildung 3.1c zu sehen. Dabei werden die Strukturen in beiden Ansichten übereinander gelegt gezeichnet. Geometrische Abweichungen zwischen den Strukturen werden auch im Zweidimensiona-

len durch eine Verschiebung der Strukturen veranschaulicht. Die Auswahl einer bestimmten Struktur wird in beiden Ansichten hervorgehoben dargestellt.

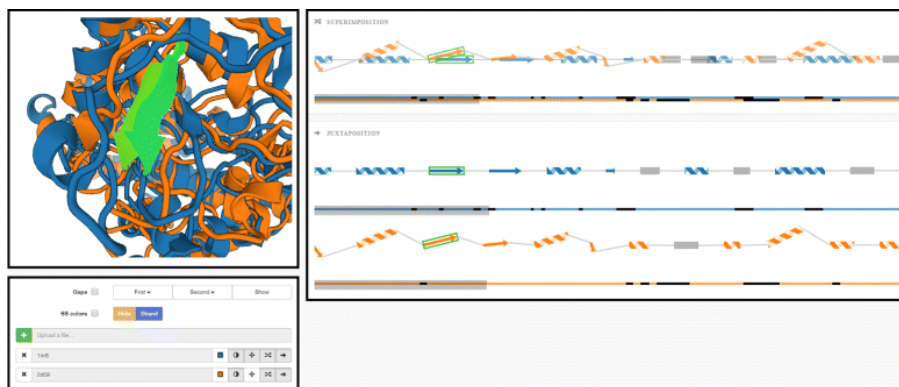
Sarikaya et al. [SAMG14] entwickelten eine Darstellung zur Validierung von Klassifikatoren, welche Moleküloberflächen zugewiesen werden. In der zweidimensionalen Darstellung stehen verschiedene Glyphen-Repräsentationen in Form eines Histogramms, einer Wahrheitsmatrix, einer Heatmap oder einem Cluster-Plot zu Verfügung. Auf der Moleküloberfläche in der dreidimensionalen Darstellung können bestimmte Bereiche entsprechend eingefärbt werden, siehe Abbildung 3.1a.

Bei Telea und Auber [TA08] wird die strukturelle Evolution von Programm-Code mit der dafür entwickelten *Code-Flow*-Visualisierung dargestellt. Dabei werden übereinstimmende Code-Zeilen in unterschiedlichen Programm-Versionen an gespiegelten, sogenannten Eiszapfen-Plots, durch texturierte Spline-Kurven miteinander verbunden, siehe Abbildung 3.1b.



(a) Klassifikation von Moleküloberflächen.
Quelle: [SAMG14]

(b) Evolution von Programm-Code.
Quelle: [TA08]



(c) Vergleich von Strukturunterschieden. Quelle: [KJB+17]

Abbildung 3.1: Unterschiedliche Struktur- und Sequenz-Visualisierungen.

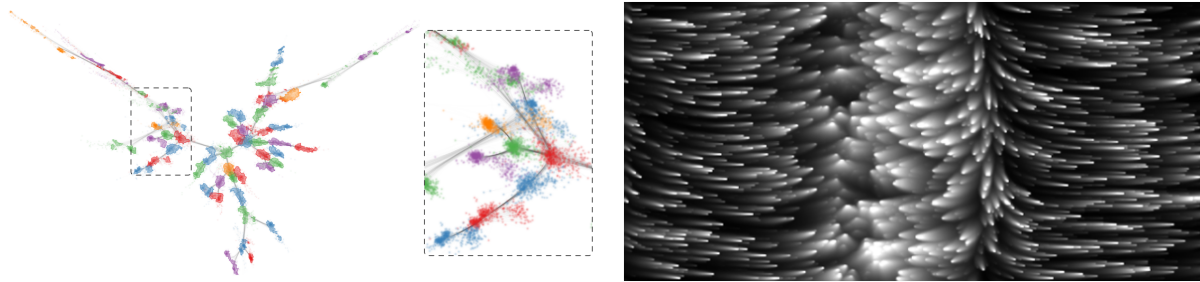
3.3 Visualisierung von Unsicherheit

Sowohl die Betonung, wie wichtig es ist, Unsicherheit von Daten zu visualisieren, als auch die Untersuchung geeigneter Methoden, Unsicherheit zu visualisieren, geht auf Arbeiten von MacEachren [Mac92], Pang, Wittenbrink und Lodha [PWL97] und Gershon [Ger98] zurück. Das Ziel der Visualisierung von Unsicherheit ist es, Daten zusammen mit zusätzlichen Informationen über deren Unsicherheit darzustellen. Die Darstellungen stellen für die Analyse eine vollständigere und genauere Wiedergabe der Daten dar [PWL97]. Unsicherheit kann bei einer Analyse das Vertrauen in das Ergebnis erhöhen. Visualisierungen von Unsicherheit sind vor allem im Bereich der Geoinformationssysteme (GIS), welche sich mit der Erfassung, Bearbeitung, Analyse und Präsentation räumlicher Daten befassen, von Bedeutung [MRH05]. Dies hat Zunehmens für die Verbreitung der Visualisierung von Unsicherheit in anderen Bereichen, wie der Strömungs- oder Volumen-Darstellung, beigetragen. Die Forschung zu diesem Thema spaltet sich dabei in die beiden, im Folgenden näher erläuterten, Bereiche der theoretischen Untersuchung von Visualisierungen und der Entwicklung neuer Darstellungen.

Im erstgenannten Bereich werden die zahlreichen Möglichkeiten, die für die Visualisierung von Unsicherheit in den unterschiedlichsten Kontexten zur Verfügung stehen, anhand verschiedener Taxonomien und Topologien wie in den Arbeiten [PWL97], [THMG05], [MRH05] und [PRJ12] kategorisiert. Dabei wird beispielsweise nach der Dimension der Unsicherheit, nach der Dimension der zugrundeliegenden Daten oder nach Kategorien der Unsicherheit unterschieden. Eine weitere Unterteilung erfolgt anhand der von Bertin in [Ber83] eingeführten *visuellen Variablen*. Diese sind als manipulierbare Primitiven grafischer Zeichenmittel definiert, mit denen jegliche Information grafisch dargestellt werden kann. Unterschiedliche visuelle Variablen werden in [GHL15], [BBIF12], [CG14] oder [SZB+09] mit Benutzerstudien evaluiert. Die visuellen Variablen werden dabei in Repräsentationen von Unsicherheit unterteilt, welche auf Punkten oder Linien basieren. Eine weitere Möglichkeit der Evaluation erfolgt bei [ZC06] anhand verschiedener theoretischer Grundsätze zur Wahrnehmung. Maceachren et al. [MRO+12] evaluieren in ihrer Arbeit anhand der visuellen Semiotik.² Eine ausführliche Betrachtung der Ergebnisse unterschiedlicher Evaluationen erfolgt in Abschnitt 5.1 bei der Analyse geeigneter Visualisierungen für den hier gegebenen Kontext.

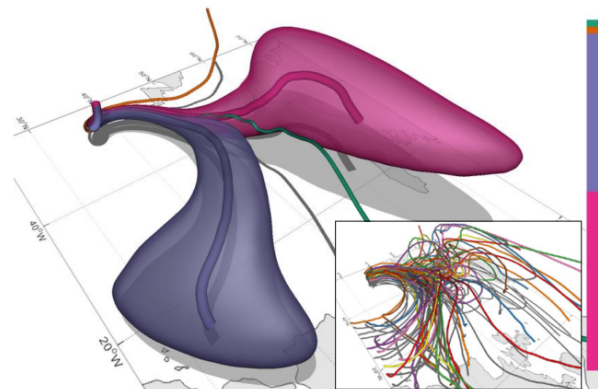
Der andere Bereich der Forschung zum Thema Visualisierung von Unsicherheit beschäftigt sich mit der Entwicklung neuer Visualisierungen für unterschiedliche Anwendungsfälle. Exemplarisch sind hier die folgenden Arbeiten, jeweils mit der verwendeten Darstellungstechnik, aufgeführt. Um statistische Eigenschaften eines Sets von Strömungslinien innerhalb eines bestimmten Bereichs eines Vektorfeldes darzustellen, verwenden Ferstl, Bürger und Westermann [FBW16] sogenannte Iso-Konturen, siehe Abbildung 3.2c. In der Arbeit von Schulz et al. [SNG+16] wird eine Methode für probabilistisches Graphen-Layout und deren Darstellung vorgestellt. Damit lassen sich unsichere Netzwerke auf ihre statistischen Eigenschaften hin

²Unter *visueller Semiotik* versteht man einen Teilbereich der Semiotik (Wissenschaft der Zeichentheorie), innerhalb welchem analysiert wird, inwiefern visuelle Bilder Informationen vermitteln.



(a) Probabilistisches Graphen-Layout.
Quelle: [SNG+16]

(b) Unsicherheit im Strömungsfeld.
Quelle: [BWE05]

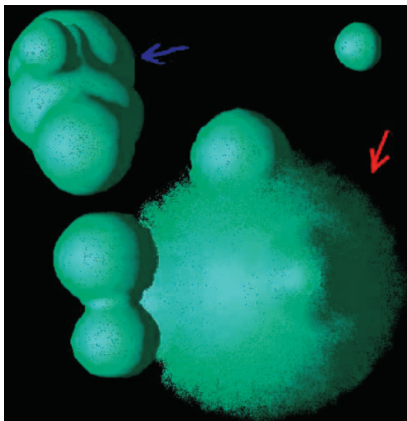


(c) Unsicherheit in Vektorfeld-Ensemble.
Quelle: [FBW16]

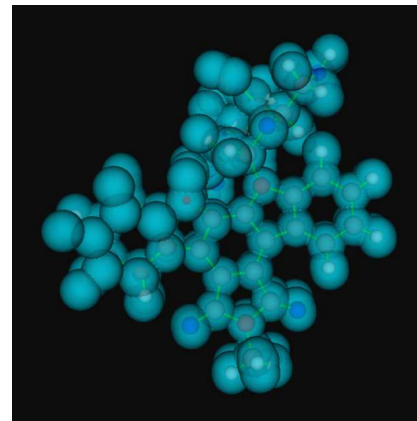
Abbildung 3.2: Verschiedene Visualisierungen von Unsicherheit.

untersuchen. Das verwendete Verfahren zur Darstellung vereint dabei die Bündelung von Kanten und Knoten sowie deren Splatting. Auch die Einfärbung von Graphen und auf der Dichte basierendes Clustering kommen zum Einsatz, siehe Abbildung 3.2a. Von Botchen, Weiskopf und Ertl [BWE05] wird eine auf Texturen basierende Visualisierung von Unsicherheit in zeitabhängigen zweidimensionalen Strömungsfeldern beschrieben. Die dafür entwickelten Darstellungskonzepte beruhen auf der Textur-Advektion, um die Richtung der Strömung durch Streichlinien dazustellen. Die Unsicherheit wird über die Unschärfe der Streichlinien vermittelt, siehe Abbildung 3.2b.

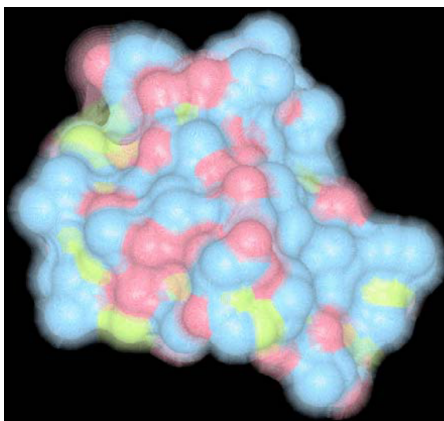
Konkret mit der Darstellung von Unsicherheit unterschiedlicher Eigenschaften bei Molekülen beschäftigen sich die Arbeiten im Weiteren. Bei der auf [GR02] basierenden Arbeit von Grigoryan und Rheingans [GR04] wird für die Darstellung der Unsicherheit von Moleküloberflächen ein auf Punkten basierender Ansatz verwendet, siehe Abbildung 3.3a. Die Unsicherheit der Lage einzelner Atome eines Moleküls wird bei Rheingans und Joshi [RJ99] über Splatting mittels Gaußglocke und der Extraktion von Iso-Flächen dargestellt, siehe Abbildung 3.3b.



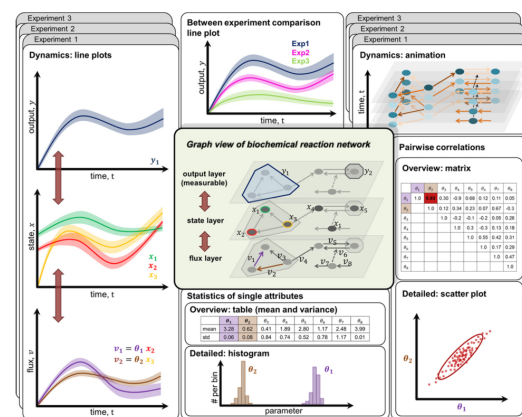
(a) Unsicherheit von Moleküloberflächen. Quelle: [GR04]



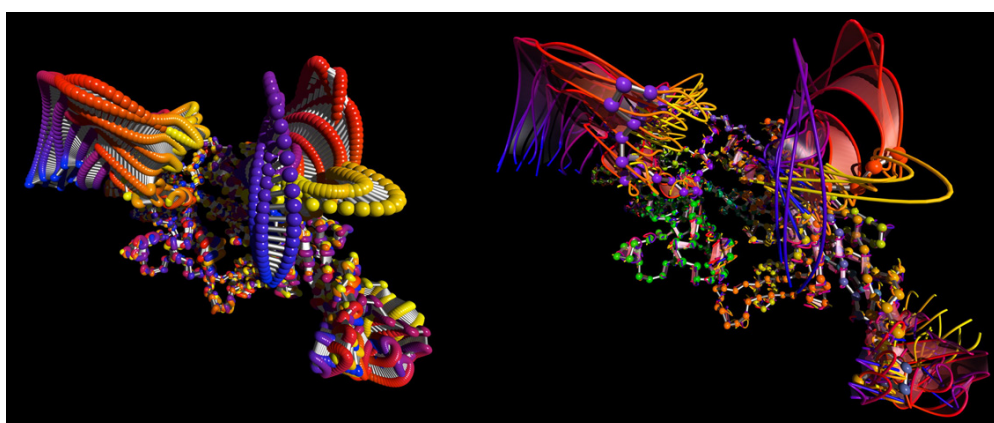
(b) Unsicherheit der Position von Atomen. Quelle: [RJ99]



(c) Unsicherheit durch thermische Vibrationen. Quelle: [LV02]



(d) Unsicherheit in Netzwerken biochemischer Reaktionen. Quelle: [VHK+13]



(e) Konformationsänderungen in Molekülen. Quelle: [DRSR15]

Abbildung 3.3: Visualisierungen von Unsicherheit bei Molekülen.

Auch bei Lee und Varshney [LV02] war es Ziel der Arbeit die Unsicherheit der Position von Atomen durch die Darstellung der thermischen Vibration zum Ausdruck zu bringen. Die verwendete Darstellungstechnik beruht auf mehrschichtigen transparenten Oberflächen, siehe Abbildung 3.3c. In der Arbeit von Vehlow et al. [VHK+13] wird die Analyse unsicherer Eigenschaften von Netzwerken biochemischer Reaktionen anhand verschiedener zweidimensionaler Diagramme ermöglicht, siehe Abbildung 3.3d. Bei Dabdoub et al. [DRSR15] werden Konformationsänderungen in Molekülen mit einer auf den Pfaden der Atome basierenden Repräsentation dargestellt, siehe Abbildung 3.3e. Eine aktuelle Zusammenfassung der Forschung im Bereich der Visualisierungen von Unsicherheit wird zum Beispiel in [JS03] oder [Riv07] gegeben.

Kapitel 4

Modell für Unsicherheit

Es gibt keine Sicherheit, nur
verschiedene Grade der Unsicherheit.

— Anton Pawlowitsch Tschechow

Im Folgenden Kapitel wird das Modell zur Bestimmung der Unsicherheit aufgrund unterschiedlicher Sekundärstrukturzuordnungen beschrieben. Es wird entsprechend der Visualisierungspipeline für Unsicherheit auf die Erfassung und Umwandlung der Daten und die sich ableitende Unsicherheit eingegangen (siehe Abbildung 2.1 auf Seite 9). Zur Formulierung eines Modells war es im ersten Schritt notwendig, die zu berücksichtigenden Quellen von Unsicherheit im Verlauf der Strukturzuordnung und bei deren Vergleich zu identifizieren. Die Hauptursache, warum die einzelnen Verfahren zum Teil stark von einander abweichende Sekundärstrukturen zuordnen, liegt darin, dass es für die Zuordnung der verschiedenen Strukturtypen keine einheitlich definierten Kriterien und somit keine Referenz gibt. Es ist somit an Stellen mit verschiedenen Strukturzuordnungen nicht eindeutig zu sagen, welche Struktur tatsächlich vorliegt. Es ist keine absolute Aussage über die Richtigkeit einer Strukturzuordnung möglich, sondern nur relativ zwischen den Verfahren, die jeweils nur im Rahmen ihrer Definitionen beurteilt werden können. Die manuelle Zuweisung wird allerdings im Allgemeinen als Referenz angesehen und bei der Beurteilung als solche verwendet. Ein weiterer Punkt bei den Verfahren ist die unterschiedlich detaillierte Differenzierung in einzelne Sekundärstrukturtypen. Davon ausgehend werden im Weiteren sowohl auf der Ebene der Verfahren als auch auf der Ebene der einzelnen Aminosäuren Kriterien zur Berechnung der Unsicherheit definiert. Die Datengrundlage des Berechnungsmodells sind dabei die Strukturzuordnungen der gebräuchlichen Verfahren STRIDE, DSSP und die in den PDB-Dateien angegebenen Sekundärstrukturen. PDB wird im Folgenden der Einfachheit halber auch als Verfahren bezeichnet, wobei dabei das jeweilige Verfahren gemeint ist, das zur Bestimmung der Sekundärstruktur innerhalb einer PDB-Datei verwendet wurde. Die Bestimmung der Sekundärstruktur in den PDB-Dateien erfolgt dabei standardmäßig, für die meisten Einträge automatisiert, durch das von *Worldwide Protein Data Base* [BHN03] in seinen *Processing Procedures* [Sta14] bevorzugte Verfahren PROMOTIF. Alternativ

dazu wird die Sekundärstruktur von DSSP oder durch manuelle Zuordnung eines Experten bzw. eines Kristallographen bestimmt, wobei es noch einen zu vernachlässigenden Anteil an weiteren Zuordnungsmöglichkeiten gibt (z.B. PROCHECK).¹ Durch den geringen Anteil an PDB-Einträgen mit von Experten manuell zugeordneten Sekundärstrukturen wird PROMOTIF als zusätzliches Verfahren bei der Unsicherheitsberechnung mit einbezogen. Allerdings sind in den PDB-Dateien nur die Sekundärstrukturen für Helices und Faltblätter angegeben. Dies sind zwar die wichtigsten Strukturtypen, sie stellen aber nur einen Teil der durch PROMOTIF bestimmten Sekundärstrukturen dar. Das Ziel des hier entwickelten Berechnungsmodells war es, sowohl eine qualitative als auch quantitative Abschätzung der Unsicherheit für den hier vorliegenden Fall zu beschreiben. Bei den Strukturtypen handelt es sich um rein qualitative Daten und nicht um konkrete numerische Messwerte, sodass deren Unsicherheit nicht absolut sondern relativ bzw. prozentual ausgedrückt wird. Auch eine quantitative Aussage erfolgt über Wahrscheinlichkeiten. Das Modell folgt also einem rein probabilistischen Ansatz. Der Standardleitfaden für die Berechnung von Unsicherheit [Jcg08] diente hierbei als Maßgabe.

4.1 Genauigkeit der Atom-Koordinaten

Die in den Verfahren beschriebenen Kriterien für die Zuordnung von Sekundärstrukturen beruhen neben der Information über die Sequenz der Aminosäuren hauptsächlich auf den in den PDB-Dateien angegebenen Koordinaten der einzelnen Atome eines Proteins. Die Genauigkeit der Atom-Koordinaten (angegeben in Ångström) im PDB-Dateiformat ist dabei durch die verwendete Darstellung der Fließkommazahlen im Format „Real(8,3)“ auf drei Stellen hinter dem Komma begrenzt.² Dabei hängt die Qualität der Strukturzuordnungen mit der Auflösungsgenauigkeit der im Experiment gewonnenen Koordinaten direkt zusammen. Denn aus einer höheren Auflösung der Atom-Koordinaten folgen unmittelbar eine besseren Topologie (räumliche Anordnung) und Stereochemie (dreidimensionaler Aufbau) des daraus resultierenden Proteinmodells [MMHT92], was somit auch die Genauigkeit der Strukturzuordnungen erhöht. Die Auflösungsgenauigkeit schwankt dabei überwiegend in einem Intervall von ungefähr 1.0

¹Eine von PROMOTIF abweichende Zuordnung der möglichen Sekundärstrukturen *HELIX* und *SHEET* muss in der PDB-Datei jeweils explizit als Vermerk (REMARK 650 bzw. 700) angegeben werden (siehe [Sta14]). Die genauen Anteile der alternativ zu PROMOTIF vorkommenden Verfahren lassen sich anhand einer Suche unter www.rcsb.org [BWF+00] mit dem Suchbegriff *DETERMINATION METHOD* bestimmen. Dies ergibt 3119 Treffer. Diese Einträge wurden auf textueller Ebene für die Ermittlung der durch DSSP bestimmten Strukturen mit *DSSP* oder *KABSCH & SANDER* durchsucht, was 1575 Treffer ergab. Dies entspricht bei insgesamt 127184 Einträgen unter rcsb.org für DSSP einem Anteil von ca. 1,2%. Für die manuelle Zuordnung ergab eine Suche nach *DEPOSITOR* oder *AUTHOR* 1368 Treffer. Dies entspricht ca. 1,1% aller Zuordnungen. Für PROMOTIF bedeutet dies einen Anteil von ca. 97,5%. Stand der Zahlen: 02.03.2017.

²Ein Ångström entspricht $10^{-10}m$ und ist eine von der typischen Größenordnung eines Atomradius abgeleitete Längeneinheit.

bis 3.5 Ångström.³ Die hauptsächlich angewendete experimentelle Methode zur Strukturanalyse von Proteinen ist die Röntgenstrahlenstrukturanalyse (siehe Abschnitt 2.1.1) mit einem Anteil von ca. 89%.⁴ Als konkurrierende Alternative zur Röntgenkristallographie wird für kleine bis mittelgroße Proteine (< 70 kDa) die Kernspinresonanzspektroskopie verwendet. Eher selten kommen noch die Elektronenmikroskopie und Hybrid-Verfahren zur Bestimmung der Atom-Koordinaten zum Einsatz.⁵

Zur Fehlerabschätzung des gewonnenen Modells wird häufig der sogenannte *R-Faktor* als globaler Indikator verwendet. Im *R-Faktor* kombiniert sind der in den experimentellen Daten enthaltene Fehler und die Abweichung des Modells von der Realität (siehe [WMDJ08] Seite 10). Verschiedene weitere Faktoren zur Strukturvalidität sind beispielsweise unter www.rcsb.org angegeben und generell gilt, dass eine Struktur für einen *R-Faktor* < 0.2 ausreichend genau ist. Eine daraus abgeleitete Unsicherheit bezüglich der Genauigkeit des Strukturmodells würde pro Protein und somit für jede darin vorkommende Aminosäure gleichermaßen gelten. Ein globaler Faktor würde zur hier angestrebten Differenzierung der pro Aminosäure auftretenden Zuordnungsunsicherheit nichts beitragen. Beim Vergleich zwischen verschiedenen Proteinen wäre es allerdings sinnvoll, anhand eines solchen Faktors die Zuordnungsunsicherheiten korrelieren zu können. Der sogenannte Temperaturfaktor oder *B-Faktor*, welcher auch in den ATOM-Einträgen der PDB-Dateien aufgeführt ist, kann als Maß für eine Abweichung pro Atom gesehen werden. Die Berücksichtigung des möglichen Einflusses von Modellfehlern auf Ebene der Atome auf die Strukturzuordnungen wäre für den Rahmen dieser Arbeit zu umfangreich. Für mehr Details zur Fehlerabschätzung siehe [Ric81] Abschnitt „C. Levels of Error“. Die wichtigste Verallgemeinerung aus Untersuchungen zur Fehlerabschätzung bei der Kristallographie ist laut Richardson [Ric81], dass Fehler relativ selten auftreten und 95% der veröffentlichten Informationen absolut korrekt sind. Trotzdem ist keine der gewonnenen Informationen immun gegen mögliche Fehler. Aufgrund der obigen Argumente wird hier die Messungenauigkeit der Atom-Koordinaten bei der Unsicherheitsberechnung der Strukturzuordnungen nicht mit einbezogen. Im Rahmen dieser Arbeit wird bei der Berechnung der Unsicherheit nur der direkte Einfluss der Zuordnungsverfahren berücksichtigt.

4.2 Datenerfassung

In diesem Abschnitt wird beschrieben, wie für die Berechnung der Unsicherheit alle verfügbaren Daten der Verfahren zusammengetragen, aufbereitet und anschließend in einer

³Ein Histogramm mit den Auflösungen der verfügbaren Einträge findet sich hier:

http://www.rcsb.org/pdb/statistics/histogram.do?mdcat=refine&mditem=ls_d_res_high&minLabel=0&maxLabel=5&numOfbars=10&name=Resolution

⁴Angaben zur Anzahl an PDB-Einträgen pro experimenteller Methode:

<http://www.rcsb.org/pdb/statistics/holdings.do>

⁵Die SI-Einheit Dalton (Da) ist eine atomare Masseneinheit und entspricht $1/12$ der Masse eines Atoms des Kohlenstoff-Isotops ^{12}C .

zusätzliche Datei gespeichert werden. Die Datenerfassung erfolgt in einem separaten Vorberechnungsschritt mit Hilfe eines Skriptes und ist somit unabhängig von der Ausführung des Hauptprogramms MegaMol. Dies bietet den Vorteil, dass die zum Teil zeitaufwendige Berechnung der Daten nur einmal erfolgen muss und auch für viele Proteine gesammelt ausgeführt werden kann. Die für das Skript gewählte Programmiersprache *Python* eignet sich hier besonders aufgrund der Plattformunabhängigkeit, umfangreicher Standardbibliotheken, und einer guten Dokumentation.⁶ Für die Berechnung der Unsicherheit ist dabei die PDB-Datei eines Proteins absolut notwendig und deren Verfügbarkeit die Minimalanforderung für die Datenerfassung. Die PDB-Dateien können über den Web-Server von www.rcsb.org bezogen werden. Die Ausgabedateien der Verfahren DSSP und STRIDE können auch über Web-Server bezogen werden. Die Ausgabedatei kann von den Web-Servern entweder, falls vorberechnet, direkt abgefragt werden oder durch den Upload einer PDB-Datei berechnet werden. Alternativ dazu können auch die jeweils frei verfügbaren Programme von STRIDE und DSSP lokal auf dem Rechner ausgeführt werden.

4.2.1 Dateiformate

Die Ausgabedateien der Verfahren enthalten ausschließlich ASCII-Zeichen und es handelt sich um normale Textdateien. Werden die Programme DSSP oder STRIDE ausgeführt, erfolgt deren Ausgabe der Ergebnisse standardmäßig an der Konsole. Im Folgenden wird davon ausgegangen, dass die Ergebnisse jeweils immer in Ausgabedateien geschrieben wurden, welche als Parameter der Programme angegeben werden können.

PDB Die Beschreibung des Aufbaus der PDB-Dateien folgt dem „Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description Version 3.30 (Nov. 21, 2012)“ welcher von *Worldwide PDB* veröffentlicht wurde.⁷ Die Zeilen einer PDB-Datei sind durch einen in den ersten sechs Spalten angegebenen Namen eines Eintragstyps erklärt. Die verschiedenen Eintragstypen folgen dabei einer festen Reihenfolge innerhalb der Datei. Sie können sich dabei jeweils über mehrere Zeilen erstrecken oder zum Teil auch mehrfach vorkommen. Hier wird nur auf die Eintragstypen eingegangen, die im Weiteren verwendet wurden:

⁶Offizielle Webseite: <https://www.python.org/>

⁷Dokumentation des Dateiformats von PDB: ftp://ftp.wwpdb.org/pub/pdb/doc/format_descriptions/Format_v33_A4.pdf oder im html-Format: <http://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html>

<i>Eintragstyp</i>	<i>Beschreibung</i>
REMARK 465:	Fehlende Aminosäuren, welche keine Einträge mit Koordinaten haben und nicht unter <i>ATOM</i> oder <i>HETATM</i> auftauchen.
REMARK 650:	Zusätzliche Informationen zu <i>HELIX</i> -Einträgen. Hier verwendet um festzustellen, ob die Bestimmung der Helices durch eine andere Methode als <i>PROMOTIF</i> erfolgte. Für die manuelle Zuordnung lautet der Eintrag z.B.: <i>DETERMINATION METHOD: AUTHOR PROVIDED</i> .
REMARK 700:	Zusätzliche Informationen zu <i>SHEET</i> -Einträgen, wie oben für Helices.
HELIX:	Diese Einträge beschreiben die Position jeder Helix innerhalb des Proteins. Die Helices sind benannt, durchnummeriert und klassifiziert. Es wird jeweils die erste und letzte zur Helix gehörende Aminosäure und die gesamte Länge angegeben.
SHEET:	Hier werden die vorkommenden Stränge beschrieben. Sie sind sowohl benannt als auch durchnummeriert. Es ist auch der Bezeichner des Falblattes angegeben, zu welchem der Strang gehört. Es wird jeweils die erste und letzte zum Strang gehörende Aminosäure angegeben.
ATOM:	Hier werden unter anderem die Koordinaten für jedes Atom der Standard-Aminosäuren aufgeführt.
HETATM:	Einträge wie <i>ATOM</i> allerdings für nicht Standard-Aminosäuren.

Die Durchnummerierung der Aminosäuren folgt einer speziellen Sequenz-Nummerierung. Dies erlaubt die nahtlose Verlinkung mit bestimmten Sequenzen anderer PDB-Dateien. Daraus folgt, dass die Nummerierung der Aminosäuren an einer beliebigen Stelle starten kann und die Zahlen zusätzlich von einem Buchstaben gefolgt sein können. Im Weiteren wird die Sequenz-Nummerierung einer Aminosäure als PDB-Index bezeichnet.

STRIDE Die Ausgabe von STRIDE ist durch 79 Zeichen lange Zeilen aufgebaut, die alle durch einen Eintragstyp in den ersten drei Spalten definiert sind.⁸ Die Anzahl der Felder pro Zeile ist auch hier immer konsistent, sodass sie automatisiert verarbeitbar sind. Relevant sind hier die Einträge vom Typ *ASG*, welche für alle Aminosäuren die detaillierte Sekundärstrukturzuordnung enthalten. Das Format dieser Zeilen ist dabei folgendermaßen:

⁸Dokumentation des Dateiformats von STRIDE: <http://webclu.bio.wzw.tum.de/stride/stride.doc>

4 Modell für Unsicherheit

<i>Spalten</i>	<i>Beschreibung</i>
6-8	Name der Aminosäure im Drei-Buchstaben-Code.
10-10	Identifikator der Protein-Kette.
12-15	PDB-Index der Aminosäure.
17-20	Fortlaufende Nummerierung der Aminosäuren beginnend mit Eins.
25-25	Ein-Buchstaben-Code der Sekundärstruktur.
27-39	Vollständige Bezeichnung der Sekundärstruktur.
43-49	ϕ -Winkel.
53-59	ψ -Winkel.
65-69	Für Lösungsmittel erreichbare Fläche der Aminosäure.

DSSP Bei der DSSP-Ausgabedatei werden nach einigen zusammenfassenden statistischen Daten zum Protein die Informationen jeder Aminosäure auch wieder zeilenweise aufgeführt.⁹ Die Zeilen folgen dabei auch einer festen Formatierung:

<i>Spalten</i>	<i>Überschrift</i>	<i>Beschreibung</i>
0-5	#	Fortlaufende Nummerierung der Aminosäuren beginnend mit Eins.
6-12	RESIDUE	PDB-Index der Aminosäure und Identifikator der Kette.
14-15	AA	Name der Aminosäure im Ein-Buchstaben-Code.
17-25	STRUCTURE	Detaillierte Beschreibung der Sekundärstruktur.
27-29	BP1	Nummer des ersten Partners einer β -Brücke.
31-33	BP2	Nummer des zweiten Partners einer β -Brücke.
35-38	ACC	Erreichbarkeit durch ein Lösungsmittel.
41-50	N-H \rightarrow O	Wasserstoffbrückenbindungen.
52-61	O \rightarrow H-N	Wasserstoffbrückenbindungen.
63-72	N-H \rightarrow O	Wasserstoffbrückenbindungen.
74-83	O \rightarrow H-N	Wasserstoffbrückenbindungen.
84-91	TCO	Spezieller Winkel der nicht in die Sturkturbestimmung mit einfließt.
92-97	KAPPA	Winkel zur Definition einer Krümmung (B).
98-103	ALPHA	Winkel zu Bestimmung der Händigkeit.
104-109	PHI	ϕ -Winkel.
110-115	PSI	ψ -Winkel.
116-122	X-CA	X-Koordinate des C_α -Atoms.
123-129	Y-CA	Y-Koordinate des C_α -Atoms.
130-136	Z-CA	Z-Koordinate des C_α -Atoms.

Mit ! oder * markierte Zeilen zeigen dabei entweder eine Unstetigkeit in der Hauptkette oder eine Unterbrechung der Aminosäuren-Kette an.

⁹Dokumentation des Dateiformats von DSSP: http://swift.cmbi.ru.nl/gv/dssp/DSSP_3.html und ergänzend mit Änderungen seit 1995 <http://swift.cmbi.ru.nl/gv/dssp/HTML/format.html>

4.2.2 Ausgabe zusätzlicher Werte bei STRIDE

Beim Verfahren STRIDE beruhen die Kriterien für die Zuordnung von α -Helices und Strängen auf empirisch festgelegten Schwellenwerten. Um später die daraus resultierende Unsicherheit bei der Zuordnung dieser Strukturtypen berücksichtigen zu können, müssen die berechneten Werte, die mit den Schwellenwerten verglichen werden, zusätzlich in die Ausgabedatei des Programms mit aufgenommen werden. Dafür wurde der frei verfügbare Quellcode des STRIDE-Programms wie hier beschrieben ergänzt.¹⁰ Auf die genaue Bedeutung und Verwendung der Schwellenwerte wird dann im Abschnitt 4.3.2 detailliert eingegangen. Damit die mit den Schwellenwerten verglichenen Werte am Ende der Sekundärstruktur-Berechnung für die Ausgabe verfügbar sind, wurde in der Datenstruktur *PROPERTY*, welche die Eigenschaften jeder Aminosäure enthält, zusätzliche Variablen vom entsprechenden Typ angelegt. In der Datei *stride.c* wurde so die Zeile 127: `float Th1, Th3, Th4, Tb1p, Tb2p, Tb1a, Tb2a;` hinzugefügt. Die Bezeichnung der Variablen lehnt sich dabei an die der Schwellenwerte in [FA95] an. Die Entsprechungen sind: $Th1 \cong T_1^\alpha$, $Th3 \cong T_3^\alpha$, $Th4 \cong T_4^\alpha$, $Tb1a \cong T_{1Antiparallel}^\beta$, $Tb2a \cong T_{2Antiparallel}^\beta$, $Tb1p \cong T_{1Parallel}^\beta$ und $Tb2p \cong T_{2Parallel}^\beta$. Die Initialisierung dieser Werte erfolgt in der Datei *rdpdb.c* in den Zeilen 115-122 mit der in der verwendeten Programmiersprache C verfügbaren Makrokonstante *INFINITY*, welche den größtmöglichen Wert einer `float`-Variablen repräsentiert. Durch eine Abweichung von diesem Wert kann später leicht nachvollzogen werden, ob ein Variablen-Wert während der Berechnung berücksichtigt wurde.

Die Zuweisung der entsprechenden Werte für die α -Helix-Kriterien erfolgt in der Datei *helix.c* durch die hinzugefügten Zeilen 40-47 innerhalb der Funktion `HELIX`. Die Werte werden den Variablen dabei vor der Überprüfung zugewiesen, damit später auch die Werte bei einer Nichterfüllung der Kriterien zur Verfügung stehen. Die mit dem Schwellenwert T_1^α zu vergleichenden Werte $t_1^\alpha(a)$ und $t_1^\alpha(a+1)$ werden immer jeweils für eine Aminosäure an Position a und für die nachfolgende Aminosäure an Position $a+1$ berechnet und jeweils der Variablen `Th1` zugewiesen. Die Werte $t_3^\alpha(a-1)$ und $t_4^\alpha(a+4)$ werden entsprechend den Variablen `Th3` bzw. `Th4` der Aminosäuren an Position $a-1$ bzw. $a+4$ übergeben.

Um die Werte der Faltblatt-Kriterien zu speichern, wurden in der Funktion `LINK` die Datei *sheet.c* um die Zeilen 215-242 ergänzt. Die Funktion `LINK` testet für zwei in Frage kommende gegenüberliegende Aminosäuren, ob ein bestimmtes Muster aus zwei Wasserstoffbrückenbindungen für das Zustandekommen eines parallelen oder antiparallelen Stranges oder einer einfachen Brücke erfüllt ist. Abhängig davon, ob der Test für einen parallelen oder antiparallelen Strang erfolgt, werden in den Eigenschaften der beiden betrachteten Aminosäuren jeweils die beide Werte für den Vergleich mit dem Schwellenwert für parallele oder antiparallele Stränge gespeichert. Da eine Aminosäure für mehrere Verbindungen zu anderen Strängen in Frage kommen und darauf getestet werden kann, wird hier immer nur das kleinste Ergebnis gespeichert, da dieses letztlich für eine Entscheidung über die Zugehörigkeit zu einem Strang ausschlaggebend ist.

¹⁰Der Quellcode des STRIDE-Programms ist unter <http://webclu.bio.wzw.tum.de/stride/stride.tar.gz> zugänglich.

Die zusätzlichen Ausgabewerte werden an jede Zeile der detaillierten Sekundärstrukturausgabe (am Zeilenanfang mit *ASG* markiert) angehängt, wofür in der Datei *stride.c* in der Zeile 71 die dort definierte maximale Breite der Ausgabe von ursprünglich 80 auf 160 Zeichen erhöht wurde: `#define OUTPUTWIDTH 160`. In der Datei *report.c* erfolgt in der Funktion `ReportDetailed` die Formatierung der Ausgabe. Diese wurde in den Zeilen 163-202 um die Spaltenbeschriftung mit den tatsächlichen Schwellenwerten und den Schwellenwertnamen ergänzt. Für die darunter folgende Auflistung der Wertes stehen dabei je Wert zehn Zeichen zur Verfügung und es werden drei Nachkommastellen angegeben.

```

1 2 3      72
REM ... |----- Secondary structure thresholds -----| 150
REM ... | -230.00 | 0.12 | 0.06 | -310.00 | -310.00 | -240.00 | -240.00 |
REM ... | ---Th1--- | ---Th3--- | ---Th4--- | ---Tb1p--- | ---Tb2p--- | ---Tb1a--- | ---Tb2a--- |
ASG ... | 0.000 | 0.000 | 0.000 | inf | inf | -2032.688 | -2453.662 |
ASG ... | 0.000 | 0.000 | 0.000 | inf | inf | -2235.583 | -2078.583 |
ASG ... | 0.000 | 0.000 | 0.000 | inf | inf | -2054.793 | -2095.119 |
ASG ... | 0.000 | 0.000 | 0.000 | inf | inf | -1956.848 | -1456.718 |
ASG ... | 0.000 | 0.279 | 0.279 | inf | inf | inf | inf |
ASG ... | 0.000 | 0.000 | 0.000 | inf | inf | inf | inf |
ASG ... | 0.000 | 0.147 | 0.147 | inf | inf | inf | inf |
ASG ... | 0.000 | 0.000 | 0.000 | inf | inf | inf | inf |
ASG ... | 0.000 | 0.000 | 0.000 | inf | inf | -808.033 | -2200.047 |
ASG ... | 0.000 | 0.000 | 0.000 | -1584.792 | -2260.150 | -2352.646 | -2113.062 |
ASG ... | 0.000 | 0.000 | 0.000 | -2330.481 | -2104.728 | -2128.991 | -2417.057 |

```

Listing 4.1: Beispielausschnitt der erweiterten Ausgabe von STRIDE (in der ersten Zeile sind die Spaltennummern angegeben).

4.2.3 Einlesen der Strukturzuordnungen

Beim Aufruf des Skriptes ist es obligatorisch, die PDB-ID im Vier-Buchstaben-Code eines Proteins anzugeben. Optional kann über den Parameter *-d* eine ausführliche Debug-Ausgabe während der Datenerfassung aktiviert werden. Die Speicherung der notwendigen Eingabedateien und der Ausgabedatei der Datenerfassung erfolgt in einem separaten *cache*-Ordner. Dieser kann in der Funktion `GenerateDataFile` durch Änderung der Variablen `CacheLocation` genauso angepasst werden wie in darauf folgenden Variablen die Namen der erzeugten Dateien. Die Ausgabedatei wird bei Aufruf immer neu erstellt und eine existierende wird überschrieben. Ist die entsprechende PDB-Datei eines Proteins nicht schon im *cache*-Ordner vorhanden, wird sie über den Web-Server von RCSB [BWF+00] heruntergeladen.¹¹ Die Verfügbarkeit der PDB-Datei ist für eine erfolgreiche Datenerfassung zwingend. Für die Generierung der Ausgabedateien von STRIDE und DSSP stehen zwei Möglichkeiten zur Wahl. Standardmäßig wird versucht, die für eine PDB-ID vorberechneten Dateien von den offiziellen Web-Servern von STRIDE [HF04] und DSSP [TBB+15] per POST Anfrage herunterzuladen.¹² Bei gescheiterter Web-Abfrage

¹¹Direkter Link zu PDB-Dateien: <http://files.rcsb.org/download/<PDB-ID>.pdb>

¹²Web-Server für STRIDE: <http://webclu.bio.wzw.tum.de/cgi-bin/stride/stridecgi.py>. Web-Server für DSSP: <http://www.cmbi.ru.nl/xssp/>

werden die lokal auf dem Computer vorhandenen Programme aufgerufen. Durch Angabe des Parameters `-o` beim Skriptaufruf kann die Ausführung dieser Programme erzwungen werden, ohne dass eine Web-Abfrage erfolgt. Um die Ausgabe der Schwellenwerte von STRIDE zu erhalten ist dieses Vorgehen notwendig, damit das oben modifizierte Programm verwendet werden kann. Für die Programme ist in der Variablen `ProgLocation` der Funktion `GenerateDataFile` der Pfad angegeben und in folgenden Variablen können auch die Programmnamen angepasst werden. Das Vorhandensein der Ausgabedateien von STRIDE und DSSP ist optional. Fehlen die Ausgabedateien, werden die Einträge jeweils mit Leerzeichen gefüllt. Beim Einlesen der PDB-Datei werden sowohl alle fehlenden als auch die in den ATOM- und HETATM-Einträgen aufgeführten Aminosäuren aufsteigend entsprechend ihrem PDB-Index (auch vorkommende Buchstaben werden berücksichtigt) sortiert. Dadurch entsteht eine vollständige Sequenz aller Aminosäuren des Proteins, wobei fehlende und heterogene Aminosäuren durch einen extra Eintrag markiert werden. Für jede Aminosäure wird initial eine neue Datenzeile für die Ausgabedatei erstellt. Die in der PDB-Datei definierten Sekundärstrukturen HELIX und SHEET werden in jede Zeile eingetragen, für die sie gelten, und mögliche alternative Bestimmungsmethoden dieser Sekundärstrukturen werden ermittelt. Anschließend werden die Sekundärstruktur-Daten von STRIDE und DSSP eingelesen und zeilenweise entsprechende dem PDB-Index in die jeweils vorhandenen Datenzeilen einsortiert. Dadurch sind alle für eine Aminosäure verfügbaren Informationen der drei Verfahren in einer Zeile enthalten.

4.2.4 Ausgabeformat

Das Format der Ausgabedatei orientiert sich dabei an dem der Eingabedateien und ist somit auch eine mit der ASCII-Zeichenkodierung konforme Textdatei mit der Dateiendung `.uid` (für **U**ncertainty **I**nput **D**ata, da sie MegaMol wiederum als Eingabedatei dient). Der Vorteil gegenüber einer Speicherung im Binärformat oder anderen gängigen Datenstrukturen wie beispielsweise JSON ist die sowohl visuell als auch von Computerprogrammen leicht lesbare Form.¹³ Das Einlesen kann über einfache String-Operationen erfolgen, welche in der Standard-Bibliothek der von MegaMol verwendeten Programmiersprache C++ verfügbar sind. Es müssen somit keine zusätzlichen Bibliotheken Dritter eingebunden werden. Außerdem entspricht dieses Dateiformat im Anwendungsbereich der Datenverarbeitung von Molekül-Daten der gebräuchlichen und gewohnten Struktur. Die ersten acht Spalten einer Zeile beschreiben deren Inhalt, wobei wiederum die ersten acht Zeilen jeder Ausgabedatei wie folgt definiert sind:

¹³JSON = „JavaScript Object Notation“ ist ein gängiges kompaktes Dateiformat, für mehr Informationen siehe: <http://json.org/>

<i>Zeile</i>	<i>Eintrag</i>	<i>Beschreibung</i>
1	PDB-ID	Enthält die PDB-ID.
2	DATE	Angabe des Erstelldatums der Ausgabedatei.
3	REMARK	Trennzeile.
4	METHOD	Hier werden die Verfahren angegeben, für die die unten angeführten Dateneinträge gelten. Bei PDB werden für HELIX und SHEET noch jeweils die Bestimmungsmethode angegeben.
5	INFO	Namen der Datenspalten.
6	COLUMN	Hier werden die Spaltennummern in Zehner-Schritten angegeben.
7	COLUMN	Beginnend mit Null ab der achten Spalte werden hier in Einer-Schritten die Spaltennummern aufgeführt. Die Angabe der Spaltennummern erleichtert aufgrund der großen Zeilenbreite einzelne Datenspalten zu bestimmen.
8	REMARK	Trennzeile.

Anschließend folgen die Zeilen, welche tatsächlich die Werte für jede Aminosäure auflisten. Diese Zeilen beginnen jeweils mit *DATA*. Das Ende der Datei wird über eine extra Zeile am Schluss angezeigt, welche lediglich *END* enthält.

4.3 Strukturwahrscheinlichkeiten auf Ebene der Verfahren

Im ersten Schritt der Berechnung der Unsicherheit erfolgt die Betrachtung des Zustandekommens der Strukturzuordnungen auf der Ebene der Verfahren. Die Verfahren beziehen sich als Grundlage ihrer Berechnungen jeweils in erster Linie auf die Koordinaten der Aminosäuren und zusätzlich auf die Aminosäuresequenz des Proteins. Die Verfahren beziehen diese Informationen aus den PDB-Dateien. Die Unsicherheitsberechnung ist vollständig im MegaMol-Modul *UncertaintyDataLoader* implementiert. Die im vorangehenden Abschnitt in einer Datei zusammengetragenen Informationen über die Zuordnungen der Verfahren werden dort in einem ersten Schritt eingelesen und in passenden Datenstrukturen gespeichert. Um im Weiteren tatsächlich Berechnungen durchführen zu können, ist eine Quantifizierung der ausschließlich qualitativen Aussage der Strukturzuordnungen durch Zuordnungswahrscheinlichkeiten notwendig. Somit wird an jeder Aminosäure und bei jedem Verfahren jeweils wiederum jeder Strukturtyp über eine Wahrscheinlichkeit quantifiziert und dadurch vergleichbar gemacht. Verbildlichen lässt sich dies über eine dreidimensionale Matrix, bei der die erste Dimension über die Anzahl der Aminosäuren bestimmt ist. Die zweite Dimension enthält Einträge entsprechend der Anzahl der Verfahren. In der dritten Dimension erfolgt die Unterteilung anhand der unterschiedlichen Strukturtypen. Die Werte dieser Matrix geben dann die Wahrscheinlichkeiten an. Ziel dieses ersten Schrittes ist es, diese für jedes Verfahren zu initialisieren. Außerdem erfolgt für STRIDE eine Modifikation der Strukturwahrscheinlichkeiten unter Einbeziehung der Zuordnungsunsicherheit. Diese entsteht bei Zuordnungskriterien, welche durch Schwellenwerte definiert werden.

Für die weiteren Betrachtungen werden dafür folgende Mengen und Funktionen definiert: Für die Anzahl aller Aminosäuren $\|A\|$ wird die Menge der Aminosäuren-Positionen definiert: $a \in \mathbb{A} = \{0, \dots, \|A\|\} \in \mathbb{N}$. Die Menge der detaillierten Sekundärstrukturtypen im Ein-Buchstaben-Code folgt dabei der Konvention von DSSP, welcher auch STRIDE folgt: $\mathbb{S} = \{H, G, I, E, B, T, S, C\}$. Es gilt: $H \hat{=} \alpha$ -Helix, $G \hat{=} 3_{10}$ -Helix, $I \hat{=} \pi$ -Helix, $E \hat{=} (\text{Erweiterter})$ Strang $\hat{=} \text{Faltblatt}$, $B \hat{=} \beta$ -Brücke, $T \hat{=} \text{Umdrehung}$, $S \hat{=} \text{Biegung}$, $C \hat{=} \text{Zufällige Windung oder Krümmung}$. Für Strukturtypen $s \in \mathbb{S}$ können zusätzliche Indices angegeben sein: $s_{[v][a]}$. Der erste Index bezieht sich darauf, dass es sich um den Strukturtyp eines bestimmten Verfahrens $v \in \mathbb{V}$ handelt. Der zweite Index gibt den Strukturtyp an einer beliebigen aber festen Aminosäure der Position $a \in \mathbb{A}$ an. Die Menge der unterschiedlichen Zuordnungsverfahren ergibt sich zu: $\mathbb{V} = \{DSSP, STRIDE, AUTHOR_{PDB}, PROMOTIF_{PDB}\}$. Für das Wertintervall der Wahrscheinlichkeiten wird $\mathbb{P} = [0, 1] \in \mathbb{R}$ definiert. Die Funktion p gibt für einen Sekundärstrukturtyp $s_{[v][a]} \in \mathbb{S}$ die Zuordnungswahrscheinlichkeit an: $p_V(s_{[v][a]}) \in \mathbb{P}$. Die Verfahren unterscheiden dabei unterschiedlich detailliert zwischen verschiedenen Struk-

Sekundärstrukturen		PDB	DSSP	STRIDE
		PROMOTIF,AUTHOR		
I	π -Helix (5-Turn)	x	x	x
H	α -Helix (4-Turn)	x	x	x
G	3_{10} -Helix (3-Turn)	x	x	x
T	Umdrehung (Turn)		x	x
S	Biegung (Bend)		x	
C	Zufällige Windung oder Krümmung (Coil, Loop)	(x)	(x)	x
B	β -Brücke (Bridge)		x	x
E	(Erweiterter) β -Strang (Strand)	x	x	x

Tabelle 4.1: Die jeweils von den Verfahren unterschiedenen Sekundärstrukturtypen.

turtypen, wie Abschnitt 4.3 zeigt. Bei STRIDE werden Einträge, auf die ansonsten keine andere Strukturzuordnung erfolgte, mit C markiert. Bei PDB und DSSP hingegen werden Sekundärstruktur-Einträge von Aminosäuren, auf die keine der berücksichtigten Strukturtypen passen, leer gelassen. Bei DSSP wird explizit darauf hingewiesen, dass ein Leerzeichen sowohl für einen Loop oder eine Irregularität stehen kann als auch dafür, dass keine Ausgabe erfolgte oder bei der Berechnung der Zuordnung ein Fehler auftrat. Wie auch bei andern Anwendungen (z.B. bei PDBFINDER) werden diese Stellen mit C markiert.¹⁴ Dies erlaubt die Unterscheidung zu Aminosäuren, an denen für DSSP tatsächlich kein Eintrag existiert, wie beispielsweise bei fehlenden Aminosäuren oder wenn aus anderen Gründen Aminosäuren nicht berücksichtigt wurden. Bei PDB wird ebenfalls allen Aminosäuren, welche keinen Sekundärstruktur-Eintrag haben, C zugewiesen. Da die Bezeichnung C somit im Einzelfall nicht

¹⁴Siehe: <http://swift.cmbi.ru.nl/gv/pdbfinder/>

im Sinne der Autoren sein könnte, soll hier die Bedeutung dieses Eintrags um die vorangehend genannten Bedeutungen erweitert verstanden sein.

4.3.1 Initialisierung der Wahrscheinlichkeiten

Bei der Initialisierung der Strukturtyp-Wahrscheinlichkeiten werden die jeweils von den Verfahren bestimmten Zuordnungen betrachtet. Diese Sekundärstrukturtypen, hier als $s_{[v][a]}^{init}$ bezeichnet, werden jeweils mit einer Wahrscheinlichkeit von 100% initialisiert:

$$\forall v \in \mathbb{V} \text{ und } \forall a \in \mathbb{A} \text{ setze: } p_V \left(s_{[v][a]}^{init} \right) = 1$$

Bei allen anderen Strukturtypen wird die Wahrscheinlichkeit auf 0% gesetzt:

$$\forall v \in \mathbb{V} \text{ und } \forall a \in \mathbb{A} \text{ gilt } \forall s'_{[v][a]} \neq s_{[v][a]}^{init}: p_V \left(s'_{[v][a]} \right) = 0$$

Dies gilt, da die Verfahren Sekundärstrukturen unabhängig voneinander zuordnen. Auch der Einfluss zusätzlichen Wissen über das Zustandekommen der Zuordnungen wird dabei vorläufig nicht berücksichtigt. Die Initialisierung bringt damit die von vielen Autoren zu diesem Thema geteilte Meinung zum Ausdruck, dass die Verfahren, im Rahmen der von ihnen definierten Kriterien, jeweils korrekt sind (siehe z.B. [KS83] und [FA95]).

4.3.2 Unsicherheitskriterium für STRIDE

STRIDE verwendet für die Zuordnung von α -Helices, β -Strängen und β -Brücken Kriterien, die anhand empirisch festgelegter Schwellenwerte festgelegt werden. Diese wurden anhand einer erschöpfenden Suche mit plausiblen Werten optimiert. Dabei wurden jeweils die automatischen Zuordnungen durch STRIDE mit manuell zugeordneten Sekundärstrukturen verglichen. Daraufhin wurden die Schwellenwerte mit der größten Übereinstimmung gewählt [FA95]. Die Optimierung erfolgte anhand von 226 speziell ausgewählten Proteinen, für welche eine manuelle Zuordnung vorgenommen wurde. Diese bilden allerdings nur eine kleine Teilmenge aller verfügbaren Proteinstrukturen. Daraus ergibt sich der Umstand, dass die Schwellenwerte nicht generell optimal sein können und somit die Genauigkeit des Verfahrens beeinträchtigen. In Folge dessen wiederum resultiert eine Unsicherheit für die Strukturzuordnungen von STRIDE. Diese Unsicherheit soll durch das im Folgenden beschriebene Kriterium dadurch abgeschätzt werden, dass die Strukturtyp-Wahrscheinlichkeiten ausgehend von der obigen Initialisierung an Stellen, an denen die Schwellenwert-Kriterien Einfluss auf die Zuordnung haben, modifiziert werden. Ausgangspunkt sind die konkreten berechneten Werte, welche mit den Schwellenwerten jeweils verglichen werden. Dabei gilt es zum einen diejenigen Fälle zu unterscheiden, in denen das Schwellenwert-Kriterium für einen Wert erfüllt ist, der Werte aber entsprechend dicht am Schwellenwert liegt. Somit erhöht sich die Unsicherheit der Entscheidung. Zm anderen ergeben sich analog dazu die Fälle, in denen die Werte zu einer Nichterfüllung der Schwellenwertkriterien führen, jedoch dicht am Schwellenwert liegen und somit diese Entscheidung auch wiederum unsicherer wird. Dafür wurden um die jeweiligen Schwellenwerte sinnvolle Intervalle definiert, innerhalb derer sich eine Modifikation der Strukturtypen-Wahrscheinlichkeiten

mit einer veränderten Sicherheit der Entscheidung korrelieren lassen. Die konkrete Verteilung der im Einzelnen berechneten Werte wäre nur mit erheblichem Aufwand zu ermitteln. Daher wird angenommen, dass es sich bei den Verteilungen der Werte jeweils für solche, die ein Schwellenwert-Kriterium erfüllen bzw. nicht erfüllen, um Normalverteilungen handelt. Dies lässt sich damit begründen, dass es für beide Normalverteilungen einen optimalen Wert - entweder für das Zutreffen oder für die Ablehnung eines Zuordnungskriteriums - geben muss, um welchen sich die anderen Werte streuen. Die Schwellenwerte liegen also zwischen diesen Optima und wurden so bestimmt, dass für möglichst viele Werte jeweils die korrekte Entscheidung getroffen wird. Da die Entscheidung durch die Schwellenwert-Kriterien, wie oben begründet, nicht immer korrekt sein kann, wird für die Schwellenwerte ein üblicherweise verwendetes Vertrauensniveau von $p_T = 95\%$ angenommen. Das heißt, dass für 95% aller Werte die Entscheidung anhand der Schwellenwerte korrekt ist. Diese Abschätzung folgt dem empfohlenen Vorgehen in [Jcg08] im Falle, dass wie hier in Bezug auf die Unsicherheit der Werte keine weiteren Informationen verfügbar sind und dort als „Typ B Standard-Unsicherheit“ bezeichnete wird. Es werden bei STRIDE folgende Schwellenwerte definiert, welche hier in einer Menge

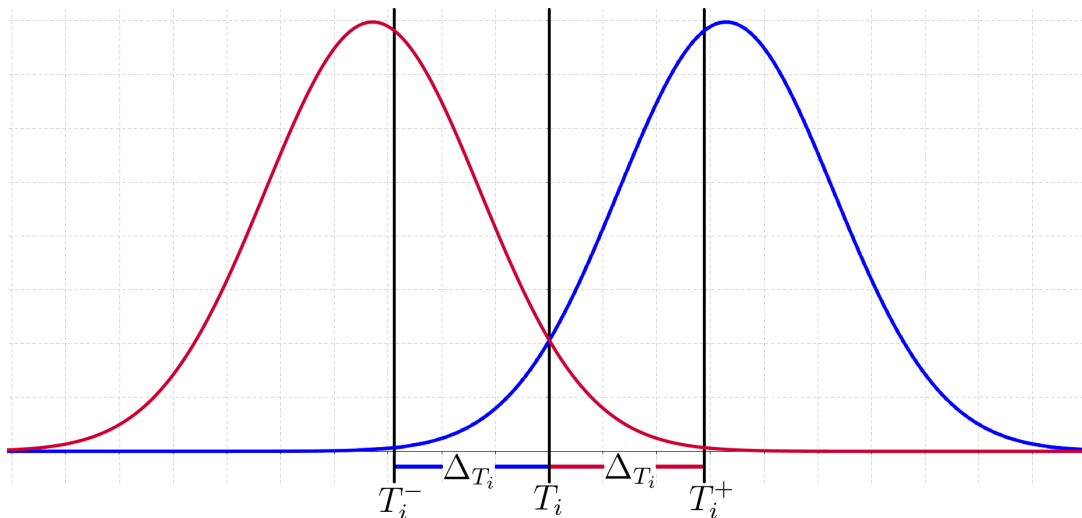


Abbildung 4.1: Normalverteilungen für Werte, welche ein Kriterium erfüllen bzw. nicht erfüllen, und das resultierende Intervall $[T_i^-, T_i^+]$ um den Schwellenwert T_i .

zusammengefasst werden: $\mathbb{T} = \{T_1^\alpha, T_3^\alpha, T_4^\alpha, T_1^\beta, T_2^\beta, T_1^\beta, T_2^\beta\}$ mit $T_i \in \mathbb{T}, \forall i \in \{1, \dots, \|\mathbb{T}\|\}$. Auf die genaue Verwendung und die konkreten Kriterien wird unten im Detail eingegangen. Bei STRIDE beeinflussen auch empirisch festgelegte Gewichte die Zuordnung, welche allerdings bei der Berechnung der Werte schon mit eingeflossen sind und somit hier nicht gesondert betrachtet werden müssen. Im Weiteren sei $v = STRIDE$ fest. Mit $t_{[v][a]}^i$ wird nachstehend der berechnete Wert bezeichnet, welcher mit dem entsprechenden Schwellenwert T_i zu vergleichenden ist. Die Abgrenzung durch einen Schwellenwert erfolgt für Werte, welche das Kriterium erfüllen bzw. nicht erfüllen, nur in eine Richtung, weshalb ein sogenannter einseitiger Vertrauensbereich gilt. Zusammen mit dem oben festgelegten

4 Modell für Unsicherheit

Vertrauensniveau von $p_T = 95\%$ und mit einer genügend großen Anzahl an Werten, für welche die Abschätzung hier gemacht wird, ergibt sich bei einer Normalverteilung ein sogenannter student-Faktor von $f_t = 1,645$ [Geo09]. Dieser gibt an, für welches Vielfache der (hier prozentual berechneten) Standardabweichung σ_{p_T} ein bestimmtes Vertrauensniveau gilt: $p_T = \sigma_{p_T} \cdot f_t$. Für die Schwellenwerte ist die Standardabweichung in Prozent somit:

$$\sigma_{p_T} = \frac{p_T}{f_t} = \frac{0,95}{1,645} \approx 0,5775 \hat{=} 57,75\%$$

Für die Definition der Intervalle um die Schwellenwerte werden die obere und untere Grenze mit T_i^+ und T_i^- beschrieben. Da eine Normalverteilung für $x \mapsto \pm\infty > 0$ ist, werden die Normalverteilungen um die beiden Optima für das Zutreffen oder Nicht-Zutreffen eines Schwellenwertkriteriums bei $p^\pm = 99,9\%$ begrenzt. Es gilt dann $f_t^\pm = 3,090$. Die Grenzen der Intervalle berechnen sich damit zu:

$$T_i^\pm = T_i \pm \left((T_i \cdot \sigma_{p_T} \cdot f_t^\pm) - T_i \right) \quad (4.1)$$

In Tabelle 4.2 sind die sich daraus ergebenden konkreten Werte für die einzelnen Schwellenwerte jeweils angeführt. Damit wird ein symmetrischer Bereich um die Schwellenwerte herum definiert, innerhalb welchem eine Veränderung der Wahrscheinlichkeit erfolgt (siehe Abbildung 4.1).

Schwellenwerte	T_i^-	T_i	T_i^+	$\Delta_{T_i} = T_i - T_i^\pm $
T_1^α	-410,435	-230,000	-49,565	180,435
T_3^α	0,026	0,120	0,214	0,094
T_4^α	0,013	0,060	0,107	0,047
$T_{1\backslash\text{antiparallel}}^\beta = T_{2\backslash\text{antiparallel}}^\beta$	-428,279	-240,000	-51,721	188,279
$T_{1\backslash\text{parallel}}^\beta = T_{2\backslash\text{parallel}}^\beta$	-553,195	-310,000	-66,805	243,195

Tabelle 4.2: Schwellenwert-Intervalle für ein einseitiges Vertrauensniveau von $p_T = 95\%$.

Die maximale Reduktion bzw. Anhebung der Wahrscheinlichkeiten für Strukturtypen, deren Werte innerhalb eines Intervalls liegen, wird mit $p_\Delta = 50\%$ festgelegt. Das heißt in den Intervallen $[T_i^-, T_i]$ bzw. $[T_i, T_i^+]$ werden dann die Wahrscheinlichkeiten der Strukturtypen in den Bereichen $[100\%, 50\%]$ bzw. $[50\%, 100\%]$ modifiziert. Da STRIDE immer eine Zuordnung vornimmt, muss sich die Wahrscheinlichkeit für eine andere Struktur in gleichem Maße erhöhen, damit die Summe aller Strukturtyp-Wahrscheinlichkeiten an jeder Aminosäure gleich Eins ist.

Es muss für ein $t_{[v][a]}^i$ gelten: $t_{[v][a]}^i < T_i$ oder $t_{[v][a]}^i \geq T_i$ und zusätzlich $|t_{[v][a]}^i - T_i| \leq \Delta_{T_i} = |T_i - T_i^\pm|$. Die Wahrscheinlichkeit einer von STRIDE an einer Aminosäure mit Position a zugewiesenen Struktur ist $p_V(s_{[v][a]}) = 1$. Diese Wahrscheinlichkeit wird dann folgendermaßen

reduziert:

$$p_V(s_{[v][a]}) = 1 - p_\Delta \cdot \left(1 - \frac{|t_{[v][a]}^i - T_i|}{\Delta_{T_i}}\right) \quad (4.2)$$

Die Erhöhung der Wahrscheinlichkeit eines anderen Strukturtyps $s'_{v_a} \neq s_{v_a}$, für den bisher $p_V(s'_{v_a}) = 0$ gilt, berechnet sich dann zu:

$$p_V(s'_{[v][a]}) = p_\Delta \cdot \left(1 - \frac{|t_{[v][a]}^i - T_i|}{\Delta_{T_i}}\right) \quad (4.3)$$

Im Folgenden wird nun beschrieben, wie sich die Schwellenwert-Kriterien für α -Helices und β -Stränge bzw. β -Brücken konkret berechnen und welche Strukturtyp-Wahrscheinlichkeiten sich unter welchen Bedingungen ändern. Da sich die Strukturen der beiden Kriterien geometrisch gegenseitig ausschließen, muss für jede Aminosäure immer nur eines der beiden Kriterien überprüft werden. Die Änderungen beeinflussen sich somit nicht wechselseitig.

α -Helix-Kriterium

Eine minimale α -Helix wird bei STRIDE dadurch definiert, dass zwischen den Aminosäure-Paaren $(a, a + 4)$ und $(a + 1, a + 5)$ zwei aufeinanderfolgende Wasserstoff-Brückenbindungen vorkommen, für die zusätzlich jeweils gilt:

$$E_{hb}^{a', a'+4} \left(1 + W_1^\alpha + W_2^\alpha \cdot \frac{P_{a'}^\alpha + P_{a'+4}^\alpha}{2}\right) = t_{[v][a']}^1 < T_1^\alpha \quad (4.4)$$

E_{hb} steht dabei für die zuvor berechnete Energie der Wasserstoff-Brückenbindung, W_1^α und W_2^α sind noch zusätzliche empirisch bestimmte Gewichte und $P_{a'+4}^\alpha$ und $P_{a'}^\alpha$ sind aus den Torsionswinkeln ϕ und ψ berechnete Wahrscheinlichkeiten. Sind die obigen Bedingungen erfüllt, wird den inneren vier Aminosäuren $a + 1, a + 2, a + 3$ und $a + 4$ der Strukturtyp H für α -Helix zugewiesen. Die Aminosäuren a bzw. $a + 5$ an den Rändern werden eingeschlossen, wenn jeweils die Bedingung $P_a^\alpha = t_{[v][a]}^3 > T_3^\alpha$ bzw. $P_{a+5}^\alpha = t_{[v][a+5]}^4 > T_4^\alpha$ erfüllt ist. Für eine detailliertere Ausführung wird auf [FA95] verwiesen. Die Modifikation der Strukturtyp-Wahrscheinlichkeiten erfolgt nun iterativ für jede Aminosäure an einer beliebigen aber festen Position a anhand der hier beschriebenen Regeln.

Beim Test, wie sicher das Hauptkriterium aus Gleichung (4.4) erfüllt ist, muss sich für zwei aufeinanderfolgende Aminosäuren eine Struktur-Wahrscheinlichkeit für α -Helix größer Null ergeben. Somit muss für die aktuell betrachtete Aminosäure a und deren Nachfolger $a + 1$ gelten: $t_{[v][a]}^1 \leq (T_1^\alpha + \Delta_{T_1^\alpha})$ und $t_{[v][a+1]}^1 \leq (T_1^\alpha + \Delta_{T_1^\alpha})$. Die Indices der beiden Aminosäuren werden für eine verkürzte Schreibweise zu $b \in \{a, a + 1\}$ zusammengefasst. Gilt für eine der Aminosäuren $t_{[v][b]}^1 < (T_1^\alpha - \Delta_{T_1^\alpha})$, wird für sie in einem Zwischenschritt $p'_V(H_{[v][b]}) = 1$ definiert. Gilt dagegen für eine Aminosäure aus b : $(T_1^\alpha - \Delta_{T_1^\alpha}) \leq t_{[v][b]}^1 < T_1^\alpha$, wird für sie

$p'_V(H_{[v][b]})$ nach Gleichung (4.2) bestimmt. Gilt im dritten Fall für eine der beiden Aminosäuren $T_1^\alpha \leq t_{[v][b]}^1 \leq (T_1^\alpha + \Delta_{T_1^\alpha})$, dann wird dafür $p'_V(H_{[v][b]})$ entsprechend Gleichung (4.3) gesetzt. Für die Aminosäuren an den Positionen $a' \in \{a, a+1, a+2, a+3\}$ werden nun die Strukturtyp-Wahrscheinlichkeiten wie folgt zugewiesen:

$$p_V(H_{[v][a']}) = \max(p_V(H_{[v][a]}), \min(p'_V(H_{[v][a]}), p'_V(H_{[v][a+1]}))) \quad (4.5)$$

Das heißt von den beiden Zwischenergebnissen entspricht die Wahrscheinlichkeit, dass das Kriterium für eine α -Helix erfüllt ist, dem niedrigeren der beiden Werte. Dabei wird dieser Wert nur dann zugewiesen, wenn für eine Aminosäure aus a' die Wahrscheinlichkeit für eine Helix vorher nicht schon höher ist. Im Gegenzug dazu müssen anschließend die Wahrscheinlichkeiten für einen anderen Strukturtypen auch entsprechend angepasst werden. Gilt nach der obigen Zuweisung an den Positionen a' je für $\forall s'_{[v][a']} \neq H : p_V(s'_{[v][a']}) = 0$, wird als alternative Struktur C gesetzt und es gilt $s'_{[v][a']} = C$. Anderenfalls wird es ein $s'_{[v][a']} \neq H$ geben, für das $p_V(s'_{[v][a']}) > 0$. Setze für das so definierte $s'_{[v][a']}$:

$$p_V(s'_{[v][a']}) = 1 - p_V(H_{[v][a']}) \quad (4.6)$$

Anschließend erfolgt für die am Rand einer Helix liegenden Aminosäuren eventuell eine Modifikation der Strukturwahrscheinlichkeiten. Für die aktuell betrachtete Aminosäure a werden die Aminosäuren mit den Indices $b \in \{a-1, a+4\}$ in Bezug auf Anfangs- oder End-Kriterium getestet. Es gelten die entsprechenden Schwellenwerte T_i^α mit $i \in \{3, 4\}$. Zusätzlich gelte für die Aminosäuren jeweils $a' \in \{a, a+1, a+2, a+3, \}$: $p_V(H_{[v][a']}) > 0$. Wenn dann weiterhin $t_{[v][b]}^i \leq (T_i - \Delta_{T_i})$ gilt wird abgebrochen, da die Wahrscheinlichkeit einer α -Helix an dieser Stelle durch das Randkriterium nicht erhöht wird. Gilt $t_{[v][b]}^i \geq (T_i + \Delta_{T_i})$, wird für die Aminosäuren am Anfang oder Ende $p'_V(H_{[v][b]}) = 1$ gesetzt. Sei stattdessen in einem dritten Fall $T_i < t_{[v][b]}^i < (T_i + \Delta_{T_i})$; ein Randkriterium wäre demnach unsicher erfüllt. Nach Gleichung (4.2) wird $p'_V(H_{[v][b]})$ in einem Zwischenschritt bestimmt. Sei dagegen in einem vierten Fall $(T_i - \Delta_{T_i}) < t_{[v][b]}^i \leq T_i$; ein Randkriterium sei damit unsicher nicht erfüllt. $p'_V(H_{[v][b]})$ wird dann nach Gleichung (4.3) berechnet. In allen Fällen wird anschließend p'_V zudem mit $\min(p'_V(H_{[v][a]}), p'_V(H_{[v][a+1]}))$ multipliziert, um somit den Einfluss aus dem Hauptkriterium mit einfließen zu lassen. Die Wahrscheinlichkeit für eine zusätzliche Aminosäure mit α -Helix am Anfang oder Ende ergibt sich dann zu:

$$p_V(H_{[v][b]}) = \max(p_V(H_{[v][b]}), (p'_V(H_{[v][b]}) \cdot \min(p'_V(H_{[v][a]}), p'_V(H_{[v][a+1]})))) \quad (4.7)$$

Die Wahrscheinlichkeit für den alternativ in Frage kommenden Strukturtyp wird analog zur Definition für Gleichung (4.6) bestimmt.

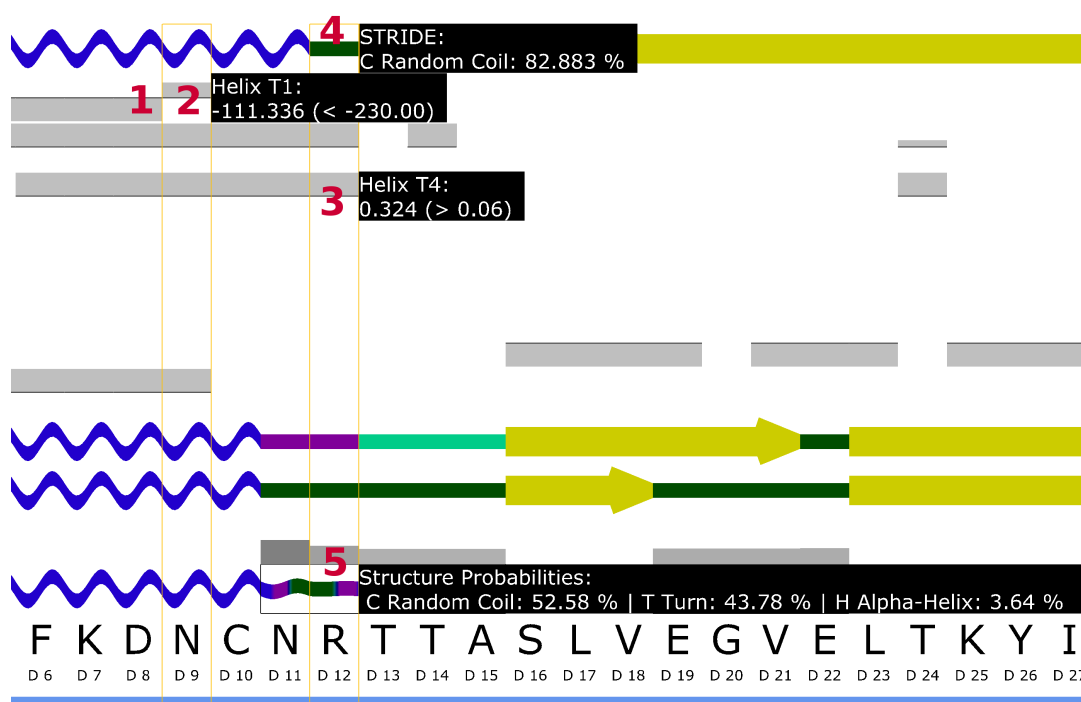


Abbildung 4.2: Beispiel, wie sich die Zuordnungsunsicherheit bei STRIDE auswirkt:

(siehe dazu auch Abschnitt 4.3.2)

1: Die Aminosäure D8 erfüllt das Kriterium für T_1^α sicher.

2: Die Aminosäure D9 erfüllt das Kriterium für T_1^α , allerdings *unsicher*.

3: Es folgt der Test für das Randkriterium an D12 für T_4^α . Dies ist sicher erfüllt.

4: Da das Kriterium an D9 unsicher erfüllt ist, resultiert eine Wahrscheinlichkeit für eine α -Helix mit 17,117%.

5: Obwohl ursprünglich keines der Verfahren eine α -Helix zugewiesen hat, ist durch die unsichere Zuordnung von STRIDE die Wahrscheinlichkeit für eine α -Helix 3,64%.

β -Strang- bzw. β -Brücken-Kriterium

Ein β -Strang setzt sich bei STRIDE aus mindestens zwei aufeinander folgenden β -Brücken zusammen. Für das Zustandekommen einer β -Brücke muss für eine Aminosäure sowohl an der Carbonylgruppe ($-CO$) als auch an der Aminogruppe ($-NH$) eine Wasserstoff-Brückenbindung ausgebildet sein. Anhängig davon, ob eine parallele oder antiparallele Anordnung der Stränge vorliegt, müssen die beiden Wasserstoffbrücken einer β -Brücke eines der nachstehenden Kriterien erfüllen:

$$E_{hb(1,2)} \left(1 + W_1^\beta + W_2^\beta \cdot CONF_{parallel} \right) < T_{1,2 \setminus parallel}^\beta \quad (4.8)$$

$$E_{hb(1,2)} \left(1 + W_1^\beta + W_2^\beta \cdot CONF_{antiparallel} \right) < T_{1,2 \setminus antiparallel}^\beta \quad (4.9)$$

$E_{hb(1,2)}$ steht dabei für die zuvor berechneten Energien der beiden Wasserstoff-Brückenbindungen. W_1^β und W_2^β sind noch zusätzliche Gewichte und $CONF_{parallel}$ bzw. $CONF_{antiparallel}$ sind Faktoren, in welchen aus den Torsionswinkeln ϕ und ψ berechnete Wahrscheinlichkeiten enthalten sind.

Die Modifikation der Strukturtyp-Wahrscheinlichkeiten erfolgt wie im Weiteren ausgeführt wieder iterativ für jede Aminosäure an einer beliebigen aber festen Position a . Dort muss sich für eine β -Brücke eine Struktur-Wahrscheinlichkeit größer Null ergeben. Im Folgenden soll jeweils für den parallelen oder antiparallelen Fall der Index $j \in \{antiparallel, parallel\}$ definiert sein. Folgende Bedingung müssen demnach für einen der beiden Fälle je für beide Wasserstoff-Brückenbindungen zutreffen: $t_{[v][a]}^{1,2\setminus j} \leq \left(T_{1,2\setminus j}^\beta + \Delta_{T_{1,2\setminus j}^\beta}\right)$. Gilt zudem weiterhin für ein festes j : $t_{[v][a]}^{1,2\setminus j} < \left(T_{1,2\setminus j}^\beta - \Delta_{T_{1,2\setminus j}^\beta}\right)$, dann wird für die Aminosäure direkt $p'_V(B_{[v][a]}) = 1$ gesetzt. Gilt dagegen: $\left(T_{1,2\setminus j}^\beta - \Delta_{T_{1,2\setminus j}^\beta}\right) \leq t_{[v][a]}^{1,2\setminus j} < T_{1,2\setminus j}^\beta$ wird $p'_V(B_{[v][a]})$ nach Gleichung (4.2) berechnet. Ist in der dritten Möglichkeit $T_{1,2\setminus j}^\beta \leq t_{[v][a]}^{1,2\setminus j} \leq \left(T_{1,2\setminus j}^\beta + \Delta_{T_{1,2\setminus j}^\beta}\right)$, dann wird $p'_V(B_{[v][a]})$ entsprechend Gleichung (4.3) festgelegt. Bei der Berechnung von $p'_V(B_{[v][a]})$ jeweils für beide Wasserstoff-Brückenbindungen, wird in allen Fällen $p'_V(B_{[v][a]})$ jeweils der kleinere Wert zugewiesen. Abhängig davon, ob an der aktuell betrachteten Aminosäure die Wahrscheinlichkeit für einen β -Strang oder eine β -Brücke größer Null ist, berechnet sich hier die resultierende Wahrscheinlichkeit ähnlich zu Gleichung (4.5). Dabei wird erst mal angenommen, dass es sich an der Stelle um eine β -Brücke handelt. Ob es sich tatsächlich um einen β -Strang handelt, wird weiter unten geprüft.

$$p_V(B_{[v][a]}) = \max\left(\max\left(p_V(B_{[v][a]}), p_V(E_{[v][a]})\right), p'_V(B_{[v][a]})\right) \quad (4.10)$$

$$p_V(E_{[v][a]}) = 0 \quad (4.11)$$

In Folge der obigen Veränderung muss anschließend noch die Wahrscheinlichkeit für einen anderen Strukturtypen angepasst werden. Gilt nach der obigen Zuweisung für $a: \forall s'_{[v][a]} \neq B: p_V(s'_{[v][a]}) = 0$ wird als alternative Struktur C gesetzt und es gilt $s'_{[v][a]} = C$. Anderenfalls wird es ein $s'_{[v][a]} \neq B$ geben, für das $p_V(s'_{[v][a]}) > 0$ ist. Setze für das so definierte $s'_{[v][a]}$:

$$p_V(s'_{[v][a]}) = 1 - p_V(B_{[v][a]}) \quad (4.12)$$

Abhängig von der Umgebung wird die Strukturtyp-Wahrscheinlichkeit der β -Brücke dann der β -Strang-Struktur zugewiesen. Wenn also $p_V(E_{[v][a-1]}) > 0$ oder $p_V(E_{[v][a+1]}) > 0$, setze dann $p_V(E_{[v][a]}) = p_V(B_{[v][a]})$ und entsprechend $p_V(B_{[v][a]}) = 0$. Ist hingegen $p_V(B_{[v][a-1]}) > 0$ oder $p_V(B_{[v][a+1]}) > 0$, verändert sich die Wahrscheinlichkeit an a für B und E , wie zuvor beschrieben. Zusätzlich werden die Wahrscheinlichkeiten für B und E auf die gleiche Weise für $a - 1$ oder $a + 1$ vertauscht.

4.4 Diskrepanzkriterien auf Ebene der Aminosäuren

Ausgehend von der Betrachtung der einzelnen Verfahren im letzten Abschnitt wird hier beschrieben, wie die Strukturtyp-Wahrscheinlichkeiten der Verfahren pro Aminosäure zusammengefasst werden. Zudem wird hier diejenige Unsicherheit quantifiziert und miteinbezogen, welche sich beim Vergleich zwischen den Zuordnungen der einzelnen Verfahren jeweils pro Aminosäure ergibt. Diese Quantifizierung geschieht anhand von paarweise zwischen Strukturtypen unterschiedlicher Verfahren definierten Diskrepanzwerten. Die Diskrepanzwerte verhalten dabei relativ zueinander. Somit findet keine Beurteilung der Verfahren in „besser“ oder „schlechter“ statt, sondern zwei Verfahren werden ausschließlich in Bezug auf eine gegenseitige Abweichung in der Strukturzuordnung in Relation gesetzt. Für zwei voneinander verschiedene Verfahren $v, w \in \mathbb{V}$ mit $v \neq w$ werden dafür die Diskrepanz-Matrizen $M_{v,w}^{s,t}$ definiert. Die Anzahl der Spalten und Zeilen entsprechen der Anzahl der unterschiedlichen Strukturtypen. Die Werte dieser Matrizen geben dabei an, inwiefern zwei Strukturtypen $s, t \in \mathbb{S}$ voneinander abweichen. Bestimmt werden die Diskrepanz-Werte $m_{v,w}^{s,t}$ anhand der im Weiteren beschriebenen Kriterien. Für eine ausreichend detaillierte Differenzierung wurde für die Diskrepanz-Werte ein Bereich von $m_{v,w}^{s,t} \in \{0, \dots, 150\} \subset \mathbb{N}$ festgelegt. Der Umstand, dass nicht jedes Verfahren alle Strukturtypen unterscheidet (siehe Abschnitt 4.3), wird in den jeweiligen Spalten der Matrizen mit „-“-Einträgen markiert. Das erste Kriterium mit Einfluss auf die Diskrepanzwerte betrachtet die von den Verfahren unabhängige generelle Verschiedenheit der Strukturtypen. Die zweite Gruppe an Kriterien, welche die Diskrepanzwerte beeinflussen, beziehen die Unterschiede zwischen den von den Verfahren definierten Kriterien zur Bestimmung der Sekundärstrukturen mit ein. In diesem Schritt wird somit eine Fortpflanzung der Unsicherheit von den einzelnen Verfahren hin zur Betrachtungsebene der Aminosäuren erreicht.

4.4.1 Generelle Strukturunterschiede

Im ersten Schritt erfolgt ein von den Verfahren unabhängiger allgemeiner Vergleich der unterschiedlichen Strukturtypen. Auch wenn sich die Verfahren in der konkreten Definition und Art der Zuordnung einer Sekundärstruktur unterscheiden, gelten für die einzelnen Strukturtypen zum Teil allgemein festgelegte Eigenschaften. Deshalb erfolgt hier für die Verschiedenheit der Strukturtypen gewissermaßen eine Vorbelegung, welche dann in den folgenden Kriterien verfeinert wird. Ausschlaggebend ist an einer bestimmten Aminosäure dabei lediglich, welcher Strukturtyp ihr zugewiesen wurde. Über wie viele Aminosäuren sich eine bestimmte Struktur dabei tatsächlich erstreckt, ist bei einem Vergleich unerheblich. Die Unterscheidungsmerkmale der Sekundärstrukturen richten sich danach, ob bzw. wie viele Wasserstoff-Brückenbindungen unter welchen Bedingungen ausgebildet sind. Jede Aminosäure kann bis zu zwei Wasserstoff-Brückenbindungen eingehen, je eine entweder als Akzeptor (Wasserstoffatom der Aminogruppe $-NH$ der Aminosäure geht Bindung ein) oder als Donor (Sauerstoff-Atom der Carbonylgruppe $-CO$ geht eine Bindung ein). Außerdem lassen sich

die Aminosäuren geometrisch anhand der dafür gebräuchlichen sogenannten Dieder- bzw. Torsionswinkel (ϕ, ψ) differenzieren. Je größer also diese Winkel sind, desto gestreckter ist die Hauptkette des Proteins. Im Folgenden wird jeder Strukturtyp auf diese Eigenschaften hin untersucht. Die biochemischen Werte wurden dabei aus [BST12] übernommen. Die Angaben zu den Verfahren wurden für STRIDE aus [FA95], für DSSP aus [KS83] und für PROMOTIF aus [HT96] entnommen.

3_{10} -Helix (G): Dieser Helix-Typ wird über eine Wasserstoff-Brückenbindung zwischen Aminosäuren an den Positionen a und $a+3$ definiert. Tatsächlich zugewiesen wird dieser Helix-Typ als Sekundärstruktur, wenn diese Bedingung mindestens für zwei aufeinander folgende Aminosäuren gilt. Die Windungsweite beträgt 3 Aminosäuren. Eine Aminosäure trägt zu einer Umdrehung einen Anteil von 120° bei. Bei mehrfacher Wiederholung nähern sich die Torsionswinkel (ϕ, ψ) den Werten $(-49^\circ, -26^\circ)$. Verallgemeinert summieren sich die Winkel ψ für eine Aminosäure a und ϕ für die Nachfolgende Aminosäure $a+1$ zu ca. -75° auf.

α -Helix (H): Bilden die Aminosäuren an den Positionen a und $a+4$ eine Wasserstoff-Brückenbindung aus, definiert sich dadurch eine α -Helix. Bilden sich zumindest zwei solcher Wasserstoff-Brückenbindungen hintereinander aus, wird auch den zugehörigen Aminosäuren dieser Strukturtyp zugewiesen. Die Windungsweite beträgt hier 3,6 Aminosäuren, was einen Anteil von 100° pro Aminosäure pro Umdrehung ausmacht. Hier bilden sich die Torsionswinkel der Hauptkette (ϕ, ψ) zu $(-60^\circ, -45^\circ)$ aus. Die Winkel ψ für eine Aminosäure a und ϕ für die Aminosäure $a+1$ summieren sich hier zu ca. -105° auf.

π -Helix (I): Die Definition richtet sich hier danach, ob zwischen zwei Aminosäuren a und $a+5$ eine Wasserstoff-Brückenbindung zustande kommt. Wiederholt sich dies mindestens einmal, erhalten die beteiligten Aminosäuren diesen Strukturtyp. Hier beträgt die Windungsweite 4,1 Aminosäuren. Eine Aminosäure entspricht damit 87° einer Umdrehung. Dieser Strukturtyp bildet keine regulären sich wiederholenden Torsionswinkel aus, wobei sich die Winkel ψ einer Aminosäure a und ϕ für die Aminosäure $a+1$ zu ca. -125° aufsummieren.

Umdrehung (Turn, T): Der Strukturtyp einer Umdrehung wird bei DSSP dann zugewiesen, wenn sich eine Wasserstoff-Brückenbindung entsprechend der obigen Definitionen der Helices nur an einer Stelle ergibt und sich nicht wiederholt. So können entsprechend Umdrehungen unterschiedlicher Windungsweiten entstehen. Alternativ dazu werden bei STRIDE und PROMOTIF sogenannte β -Umdrehungen mit vier beteiligten Aminosäuren rein geometrisch anhand verschiedener Kombinationen bestimmter Torsionswinkel definiert. Dabei ist implizit auch eine Wasserstoff-Brückenbindung Voraussetzung. Außerdem fallen noch sogenannte γ -Umdrehungen unter diesen Strukturtyp. Sie umfassen mindestens drei aufeinander folgende Aminosäuren, wobei zwischen den Aminosäuren a und $a+2$ eine Wasserstoff-Brückenbindung bestehen muss. Es treten dabei die Winkel (ϕ, ψ) $(\pm 75^\circ, \mp 65^\circ)$ auf.

Biegung (Bend, S): Dieser Strukturtyps wird nur bei DSSP definiert. Eine Biegung wird dort als rein geometrischer Strukturtyp, unabhängig vom Vorhandensein einer Wasserstoff-Brückenbindung, beschrieben und soll für Stellen mit hoher Krümmung stehen. Eine Biegung wird einer Aminosäure a dann zugewiesen, wenn der Winkel zwischen den Verbindungslinien des C_α -Atoms von a jeweils zu den C_α -Atomen der Aminosäuren $a - 2$ und $a + 2$ größer als 70° ist.

β -Brücke (Bridge, B): Für zwei sich nicht überlappende Abschnitte von je drei Aminosäuren $\{a - 1, a, a + 1\}$ und $\{b - 1, b, b + 1\}$ wird abhängig von der Ausrichtung dieser Abschnitte zwischen verschiedenen parallelen oder gegenläufigen Konfigurationen für Wasserstoff-Brückenbindungen unterschieden. Ausschlaggebend ist, dass ein Muster immer aus genau zwei Wasserstoff-Brückenbindungen besteht, welche zwischen den beteiligten Aminosäuren unterschiedlich angeordnet sein können. DSSP unterscheidet vier und STRIDE sechs solcher unterschiedlicher Anordnungen.

β -Strang (Strand, E): Der Strukturtyp eines β -Stranges wird immer dann für eine Aminosäure gesetzt, wenn an mindestens zwei aufeinander folgenden Aminosäuren das Kriterium für eine β -Brücke mit der jeweils gleichen Anordnung der Wasserstoff-Brückenbindungen erfüllt ist.

Zufällige Windung oder Krümmung (Coil, C): Trifft auf eine Aminosäure keine der anderen Strukturtypen zu, wird sie als zufällige Windung bezeichnet.

Anhand der obigen Ausführungen werden die Strukturtypen nun entsprechend ihrer gegenseitigen Abweichung über Diskrepanz-Werte korreliert. Da die Verschiedenheit der Verfahren hier noch nicht berücksichtigt wird, ergibt sich für die gleichen Strukturtypen je Verfahren eine Diskrepanz von 0, was den diagonalen Einträgen der Diskrepanz-Matrix entspricht. Aufgrund der geringsten Torsionswinkel weist eine 3_{10} -Helix die am stärksten gekrümmte Struktur auf und wird daher als Ausgangspunkt für die Bestimmung der Diskrepanzwerte gewählt. Da eine Umdrehung zwar hauptsächlich aus vier Aminosäuren besteht, sich allerdings in seltenen Fällen auch in Form einer γ -Umdrehung aus nur drei Aminosäuren zusammensetzen kann (bei DSSP treten in noch selteneren Fällen Umdrehungen bestehend aus fünf Aminosäuren auf), ist dieser Strukturtyp einer 3_{10} -Helix am ähnlichsten. Der Diskrepanzwert wird auf 15 festgelegt. Da die Abweichung zur Struktur mit den nächst größeren Winkeln, einer α -Helix, geringer ist, wird die Diskrepanz auf einen Wert von 10 gesetzt. Der nächste Schritt hin zu einer flacheren π -Helix wird auf 25 gesetzt. Das Kriterium einer Biegung trifft für die bereits genannten Strukturtypen alternativ auch in den allermeisten Fällen zu, entspricht aufgrund niedrigerer Zuordnungspriorität allerdings nicht der eigentlichen Zuordnung. Die Diskrepanz zu einer Biegung wird deshalb für die bisher genannten Strukturtypen auf 5 gesetzt. Die Diskrepanz einer Biegung zur am wenigsten gekrümmten Struktur, einer zufälligen Windung, wird auf

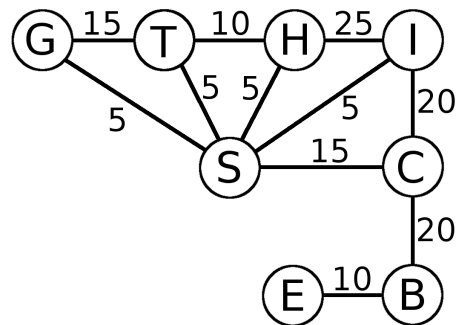


Abbildung 4.3: Veranschaulichung der Diskrepanzen zwischen den Strukturtypen. Indem die Werte auf dem längsten wiederholungsfreien Pfad zwischen zwei Strukturtypen aufsummiert werden, erhält man ihre Diskrepanz.

15 festgelegt. Entsprechend wird die Diskrepanz einer π -Helix zu einer zufälligen Windung auf 20 gesetzt. Rein geometrisch unterscheiden sich eine β -Brücke und ein β -Strang von einer zufälligen Windung nicht. Ausschlaggebend für eine Diskrepanz sind hier die Wasserstoffbrückenbindungen, welche bei diesen Strukturen vorhanden sein müssen. Die Diskrepanz zwischen einer zufälligen Windung und einer einfachen β -Brücke wird deshalb auf 20 gesetzt. Ausgehend von einer β -Brücke hin zu einem β -Strang wird die Diskrepanz dann auf 10 festgelegt. In Abbildung 4.3 sind die definierten Werte grafisch veranschaulicht. Den Diskrepanz-Wert für zwei Strukturen erhält man, indem man den längsten wiederholungsfreien Pfad von einem Strukturtyp zum anderen geht und dabei die Werte der Kanten aufsummiert. Für zwei beliebige verschiedene Verfahren * resultiert damit die nachstehende symmetrische Diskrepanz-Matrix. Die Bezeichnung der Spalten erfolgt dabei, soweit einheitlich möglich, anhand steigender Diskrepanz zwischen den Strukturtypen.

$$M_{*,*}^{s,t} = \begin{matrix} & \begin{matrix} G & T & H & I & S & C & B & E \end{matrix} \\ \begin{matrix} G \\ T \\ H \\ I \\ S \\ C \\ B \\ E \end{matrix} & \begin{pmatrix} 0 & 15 & 25 & 50 & 5 & 70 & 90 & 100 \\ 15 & 0 & 10 & 35 & 5 & 55 & 75 & 85 \\ 25 & 10 & 0 & 25 & 5 & 45 & 65 & 75 \\ 50 & 35 & 25 & 0 & 5 & 20 & 40 & 50 \\ 5 & 5 & 5 & 5 & 0 & 15 & 35 & 45 \\ 70 & 55 & 45 & 20 & 15 & 0 & 20 & 30 \\ 90 & 75 & 65 & 40 & 35 & 20 & 0 & 10 \\ 100 & 85 & 75 & 50 & 45 & 30 & 10 & 0 \end{pmatrix} \end{matrix}$$

4.4.2 STRIDE - DSSP

In diesem Abschnitt fließen in die vorangehend definierte Diskrepanz-Matrix die Abweichungen der Zuordnungskriterien zwischen STRIDE [FA95] und DSSP [KS83] ein. In der Betrachtung genereller Strukturunterschiede steckt implizit die Annahme, zwei Verfahren würden sich in

den jeweiligen Zuordnungskriterien nicht unterscheiden. Werden nun allerdings Unterschiede berücksichtigt, folgt daraus, dass sich die Diskrepanzwerte beim konkreten Vergleich von Verfahren nur erhöhen können. Da STRIDE kein Kriterium für den Strukturtyp Biegung hat, fällt diese Möglichkeit weg und die entsprechende Zeile ist durch „-“ -Einträge markiert. Aus der Sichtweise von DSSP ändert sich hingegen nichts. Die tatsächlich geänderten Diskrepanz-Werte sind in der resultierenden Matrix fett gedruckt dargestellt. Bei der Zuordnung von 3_{10} -Helices und π -Helices entspricht die Zuordnung von STRIDE der von DSSP, mit dem Unterschied, dass STRIDE eine etwas andere Definition für die Wasserstoff-Brückenbindungen verwendet. Deshalb wird die Diskrepanz zwischen diesen Strukturtypen auf 5 gesetzt. Wie bereits im letzten Abschnitt für Umdrehungen beschrieben, weichen die Definitionen bei STRIDE und DSSP hier schon deutlicher voneinander ab. Umdrehungen mit einer Windungsweite von fünf Aminosäuren kommen bei STRIDE nicht vor. Die Definition von Umdrehungen ist bei STRIDE - aufgrund der Unabhängigkeit von Helix-Kriterien - somit umfassender. Ein Vergleich von Zhang und Sagui [ZS15] ergab, dass nur ca. 40% aller von STRIDE zugewiesenen Umdrehungen auch bei DSSP einer Umdrehung entsprechen. Aufgrund dessen wird die Diskrepanz zwischen Umdrehungen auf 15 erhöht. Die Kriterien für die Zuordnung einer α -Helix unterscheiden sich bei den beiden Verfahren in der Art und Weise, wie das Vorhandensein einer Wasserstoff-Brückenbindung definiert wird. Bei DSSP gilt eine Wasserstoff-Brückenbindung als dann ausgebildet, wenn eine berechnete Bindungsenergie kleiner als $0.5 \frac{\text{kcal}}{\text{mol}}$ ist. Bei STRIDE hingegen fließen bei der Berechnung, ob eine Wasserstoff-Brückenbindung ausgebildet ist oder nicht, andere als bei DSSP berücksichtigte Energien und zusätzlich aus den Dieder-Winkeln ϕ und ψ abgeleitete Wahrscheinlichkeiten mit empirisch bestimmten Gewichte mit ein. Die Diskrepanz wird aufgrund dessen auf 10 erhöht. Wie in Abschnitt 4.3 erläutert, weist DSSP nicht explizit einen Strukturtyp zu, wenn keiner der anderen Strukturtypen zutrifft. Wie im erwähnten Abschnitt begründet, werden diese Stellen trotzdem als zufällige Windung behandelt. STRIDE weist hingegen allen Aminosäuren, welche unter kein Kriterium für einen anderen Strukturtyp fallen, konkret den einer zufälligen Windung zu. Die Bedingung, unter welcher eine zufällige Windung zugewiesen wird, ist somit bei beiden Verfahren die gleiche. Die Diskrepanz bleibt unverändert bei 0. Bei der Zuordnung von β -Brücken unterscheidet STRIDE noch zwei weitere, selten vorkommende Anordnungen für Wasserstoff-Brückenbindungen. Im Gegensatz zu DSSP werden beim Zustandekommen von β -Brücken bei STRIDE, ähnlich der Beschreibung oben bei den α -Helices, die Wasserstoff-Brückenbindungen anders definiert. Die Diskrepanz zwischen diesem Strukturtyp der beiden Verfahren wird deshalb auch auf 10 gesetzt. Liegen zwei aufeinander folgende β -Brücken gleichen Typs vor, ordnen beide Verfahren nach dem selben Kriterium einen β -Strang zu, wodurch hier keine Änderung der Diskrepanz begründet werden kann. Alle anderen im vorigen Abschnitt definierten Diskrepanz-Werte bleiben unverändert. Die konkreten Matrix-Einträge erhöhen sich allerdings durch die Aufsummierung der jeweiligen hier für einen Strukturtypen zusätzlich festgelegten Diskrepanz entsprechend. Für

den Vergleich der Strukturtypen zwischen DSSP und STRIDE ergibt sich im Ergebnis damit die folgende Diskrepanz-Matrix:

$$M_{STRIDE,DSSP}^{s,t} = \begin{matrix} & G & T & H & I & S & C & B & E \\ \begin{matrix} G \\ T \\ H \\ I \\ S \\ C \\ B \\ E \end{matrix} & \left(\begin{array}{cccccccc} \mathbf{5} & 35 & 55 & 85 & 5 & 105 & 135 & 145 \\ 35 & \mathbf{15} & 35 & 65 & 5 & 85 & 115 & 125 \\ 55 & 35 & \mathbf{10} & 40 & 5 & 60 & 90 & 100 \\ 85 & 65 & 40 & \mathbf{5} & 5 & 25 & 55 & 65 \\ - & - & - & - & - & - & - & - \\ 105 & 85 & 60 & 25 & 15 & 0 & 30 & 40 \\ 135 & 115 & 90 & 55 & 35 & 30 & \mathbf{10} & 20 \\ 145 & 125 & 100 & 65 & 45 & 40 & 20 & 0 \end{array} \right) \end{matrix}$$

4.4.3 PROMOTIF

PROMOTIF [HT96] basiert auf einer eigenen Implementierung von DSSP und folgt somit dessen Definition von Wasserstoff-Brückenbindungen. Es konnte allerdings nicht in Erfahrung gebracht werden, inwiefern sich PROMOTIF konkret von DSSP unterscheidet. Zusätzlich folgt PROMOTIF bei der Zuordnung von Helices und β -Strängen an den Rändern dieser Strukturen einer anderen Regel als STRIDE und DSSP. Die von der *Internationalen Union für reine und angewandte Chemie* (IUPAC) in [IUP74] eigentlich nur für Helices definierten Regeln 6.2 und 6.3 lassen zwei Möglichkeiten zu, wie Aminosäuren an den Rändern behandelt werden. STRIDE und DSSP folgen dabei grob der Regel 6.2, nach welcher zur Struktur diejenigen Aminosäure gehören, deren Winkel ϕ und ψ nahe denen sind, welche eine Helix definieren. Bei PROMOTIF werden entsprechend der Regel 6.3 auch diejenigen Aminosäuren dazu gezählt, welche durch Wasserstoff-Brückenbindungen an einer Struktur beteiligt sind. Wo möglich wird so bei PROMOTIF in der Praxis an den Enden von Helices und β -Strängen je eine zusätzliche Aminosäure hinzugefügt. In den PDB-Dateien werden als Strukturtypen nur Helices und β -Stränge konkret aufgeführt, wobei PROMOTIF auch noch Aussagen zu Umdrehungen macht, welche denen bei STRIDE recht ähnlich sind. Die Diskrepanz fließt allerdings bei der Aufsummierung implizit mit ein und wird in der Diskrepanz-Matrix grau dargestellt, somit bei 15 belassen. Da in PDB-Dateien auch keine β -Brücken und Biegungen berücksichtigt werden, sind die Einträge dieser drei Spalten in der Diskrepanz-Matrix je mit „-“ markiert. Beim Vergleich der Kriterien zwischen DSSP und PROMOTIF wird anhand der obigen Ausführungen für die verbleibenden Strukturtypen, mit Ausnahme der zufälligen Windungen und β -Strängen, eine Diskrepanz von 5 angenommen. Die Abweichung bei β -Brücken, welche in den PDB-Dateien

nicht angegeben sind, fließen auch indirekt mit ein. Dies ergibt die folgende Diskrepanz-Matrix:

$$M_{DSSP,PROMOTIF_{PDB}}^{s,t} = \begin{matrix} & G & T & H & I & S & C & B & E \\ \begin{matrix} G \\ T \\ H \\ I \\ S \\ C \\ B \\ E \end{matrix} & \left(\begin{array}{cccccccc} \mathbf{5} & - & 50 & 80 & - & 100 & - & 135 \\ 35 & \mathbf{15} & 30 & 60 & - & 80 & - & 115 \\ 50 & - & \mathbf{5} & 35 & - & 55 & - & 90 \\ 80 & - & 35 & \mathbf{5} & - & 25 & - & 60 \\ 5 & - & 5 & 5 & - & 15 & - & 50 \\ 100 & - & 55 & 25 & - & 0 & - & 35 \\ 125 & - & 80 & 50 & - & 25 & \mathbf{5} & 15 \\ 135 & - & 90 & 60 & - & 35 & - & 0 \end{array} \right) \end{matrix}$$

Da sich PROMOTIF und DSSP zwar unterscheiden, PROMOTIF sich vom Prinzip her allerdings sehr ähnlich wie DSSP zu STRIDE verhält, wird für die Diskrepanz zwischen PROMOTIF und STRIDE die gleiche verwendet wie schon zwischen DSSP und STRIDE. Mit Ausnahme der Umdrehungen, welche beide Verfahren sehr ähnlich definieren.

$$M_{STRIDE,PROMOTIF_{PDB}}^{s,t} = \begin{matrix} & G & T & H & I & S & C & B & E \\ \begin{matrix} G \\ T \\ H \\ I \\ S \\ C \\ B \\ E \end{matrix} & \left(\begin{array}{cccccccc} \mathbf{5} & - & 40 & 70 & - & 90 & - & 130 \\ 20 & \mathbf{0} & 20 & 50 & - & 70 & - & 110 \\ 40 & - & \mathbf{10} & 40 & - & 60 & - & 100 \\ 70 & - & 40 & \mathbf{5} & - & 25 & - & 65 \\ - & - & - & - & - & - & - & - \\ 90 & - & 60 & 25 & - & 0 & - & 40 \\ 120 & - & 90 & 55 & - & 30 & \mathbf{10} & 20 \\ 130 & - & 100 & 65 & - & 40 & - & 0 \end{array} \right) \end{matrix}$$

4.4.4 Manuelle Zuordnung

Bei der manuellen Zuordnung ordnen die Kristallographen Sekundärstrukturen anhand unterschiedlichster Kriterien zu, wie sie beispielsweise in [Fie76] beschrieben sind. Da es aber keine einheitliche Vorgehensweise gibt, unter welchen Bedingungen welche Kriterien anzuwenden sind, ist somit auch nicht ersichtlich, welche Kriterien bei einer Zuordnung konkret zur Anwendung kamen. Die Qualität einer Zuordnung ist auch von der Erfahrung eines Kristallographen abhängig. Hinzu kommt, dass in PDB-Dateien nur eine Aussage über die vorhandenen Zuordnungen möglich ist. Es ist nicht ersichtlich, ob an Stellen ohne Zuordnung tatsächlich keine Zuordnung vom Experten beabsichtigt war oder ob diese Stellen einfach nicht berücksichtigt wurden und nur deshalb keine Zuordnung erfolgte. Für die in den PDB-Dateien nicht differenzierten Strukturtypen folgen hier die gleichen wie oben bei PROMOTIF beschriebenen „-“-Einträge in den Diskrepanz-Matrizen. Stimmen die Verfahren STRIDE und DSSP mit der manuellen Zuordnung überein, wird die Diskrepanz auf 0 gesetzt. Bei verschiedenen Zuordnungen wird eine generelle Diskrepanz von 150 definiert, was letztlich zum Ausdruck bringen

soll, dass eine manuelle Zuordnung eben als bestmöglich angesehen wird. Dies entspricht somit hier der größtmöglichen Diskrepanz. Die Diskrepanz-Matrizen zwischen DSSP bzw. STRIDE und der manuellen Zuordnung ergeben sich damit wie folgt:

$$M_{DSSP,AUTHOR_{PDB}}^{s,t} = \begin{matrix} & G & T & H & I & S & C & B & E \\ \begin{matrix} G \\ T \\ H \\ I \\ S \\ C \\ B \\ E \end{matrix} & \begin{pmatrix} 0 & - & 150 & 150 & - & 150 & - & 150 \\ 150 & - & 150 & 150 & - & 150 & - & 150 \\ 150 & - & 0 & 150 & - & 150 & - & 150 \\ 150 & - & 150 & 0 & - & 150 & - & 150 \\ 150 & - & 150 & 150 & - & 150 & - & 150 \\ 150 & - & 150 & 150 & - & 0 & - & 150 \\ 150 & - & 150 & 150 & - & 150 & - & 150 \\ 150 & - & 150 & 150 & - & 150 & - & 0 \end{pmatrix} \end{matrix}$$

$$M_{STRIDE,AUTHOR_{PDB}}^{s,t} = \begin{matrix} & G & T & H & I & S & C & B & E \\ \begin{matrix} G \\ T \\ H \\ I \\ S \\ C \\ B \\ E \end{matrix} & \begin{pmatrix} 0 & - & 150 & 150 & - & 150 & - & 150 \\ 150 & - & 150 & 150 & - & 150 & - & 150 \\ 150 & - & 0 & 150 & - & 150 & - & 150 \\ 150 & - & 150 & 0 & - & 150 & - & 150 \\ - & - & - & - & - & - & - & - \\ 150 & - & 150 & 150 & - & 0 & - & 150 \\ 150 & - & 150 & 150 & - & 150 & - & 150 \\ 150 & - & 150 & 150 & - & 150 & - & 0 \end{pmatrix} \end{matrix}$$

4.4.5 Berechnung der Strukturtyp-Wahrscheinlichkeiten

Auf Grundlage der Strukturtyp-Wahrscheinlichkeiten p_V der einzelnen Verfahren aus Abschnitt 4.3 werden hier nun anhand der gegebenen Diskrepanz-Matrizen diese Wahrscheinlichkeiten je Aminosäure zusammengefasst. Im Ergebnis wird somit jedem Strukturtyp an jeder Aminosäure nur noch eine Wahrscheinlichkeit p_A zugeordnet.

Setze $M_{max} = \max(M_{v,w}^{s,t} : \forall v, w \in \mathbb{V} \text{ und } \forall s, t \in \mathbb{S})$. Da hier $M_{v,w}^{s,t} \in \{0, \dots, 150\}$ gilt, ist hier $M_{max} = 150$. In einem Zwischenschritt werden zuerst die Werte $\bar{p}(s_{[a]})$ für alle $s \in \mathbb{S}$ und für eine beliebige aber feste Aminosäure a berechnet:

$$\bar{p}(s_{[a]}) = \sum_{\forall v \in \mathbb{V}} \left(\sum_{\forall w \in \mathbb{V} \setminus v} \left(\sum_{\forall t \in \mathbb{S}} \frac{p_V(s_{[v][a]}) \cdot p_V(t_{[w][a]})}{(1 + M_{max} - M_{v,w}^{s,t})^2} \right) \right) \quad (4.13)$$

In Gleichung (4.13) wird für einen konkreten Strukturtyp $s_{[v][a]}$ bei jedem Verfahren v mit allen anderen Strukturtypen $t_{[w][a]}$ der von v verschiedenen Verfahren w der jeweilige Einfluss der Strukturtyp-Wahrscheinlichkeiten auf Ebene der Verfahren p_V mit den entsprechenden

Diskrepanzwerten aus $M_{v,w}^{s,t}$ verrechnet. Ist dabei die Wahrscheinlichkeit p_V für eine der paarweise betrachteten Strukturen gleich Null, ist auch der Summand gleich Null und liefert keinen Beitrag zum Gesamtergebnis. Das heißt umgekehrt, dass nur die Fälle relevant sind, in denen Strukturen mit einer Wahrscheinlichkeit p_V größer als Null betrachtet werden. Da die Verfahren ihre Zuordnungen unabhängig voneinander vornehmen, ergibt sich bei der gegenseitigen Berücksichtigung, dass die Wahrscheinlichkeiten miteinander multipliziert werden. Die Wahrscheinlichkeiten in $\bar{p}(s_{[a]})$ sind somit proportional zu den resultierenden Strukturtyp-Wahrscheinlichkeiten $p_A(s_{[a]})$. Die paarweisen Werte aus den Diskrepanz-Matrizen fließen hingegen negiert und umgekehrt proportional mit ein. Damit der Nenner nicht Null wird, und um den linearen Einfluss beizubehalten, liegen hier die Werte zwischen 1 und $1 + M_{max}$. Bei größer werdender Diskrepanz zwischen zwei Strukturtypen verschiedener Verfahren wird so der resultierende Nenner kleiner, und der Quotient gewinnt somit an Einfluss auf das Gesamtergebnis. Damit soll zum Ausdruck gebracht werden, dass desto unterschiedlicher die Zuordnungen der Verfahren sind, umso stärker die Unsicherheit der Sekundärstruktur an einer Aminosäure steigt. Um den Diskrepanzwerten das gleiche Gewicht wie den Wahrscheinlichkeiten zu geben, wird ihr Anteil auch quadratisch berücksichtigt. Durch Normierung der $\bar{p}(s_{[a]})$ ergeben sich die resultierenden Strukturtyp-Wahrscheinlichkeiten auf Ebene der Aminosäuren zu:

$$p_A(s_{[a]}) = \frac{\bar{p}(s_{[a]})}{\sum_{\forall s \in \mathbb{S}} \bar{p}(s_{[a]})} \in [0, 1] \quad (4.14)$$

Zudem gilt somit, dass sich die Strukturtyp-Wahrscheinlichkeiten an einer Position a zu Eins aufsummieren:

$$\sum_{\forall s \in \mathbb{S}} p_A(s_{[a]}) = 1 \quad (4.15)$$

4.5 Reduktion auf einen Unsicherheitswert

Im finalen Schritt der Unsicherheitsberechnung werden die im letzten Abschnitt für jede Aminosäure zusammengefassten Strukturtyp-Wahrscheinlichkeiten hier nun auf einen Unsicherheitswert reduziert. Im Weiteren dem Leitfaden zur Berechnung von Unsicherheit [Jcg08] folgend, werden die Strukturtyp-Wahrscheinlichkeiten je Aminosäure als eine diskrete Verteilung von Wahrscheinlichkeiten über die möglichen Werte, welche den verschiedenen Strukturtypen entsprechen, angesehen. Als gängiges Maß für die Unsicherheit wird die Standardabweichung verwendet. Definiert sei für die folgenden Betrachtungen ein Vektor über alle Strukturtyp-Wahrscheinlichkeiten für eine feste Position a :

$\vec{p}^a = (p_{s_1}^a, p_{s_2}^a, \dots, p_{s_S}^a)$ für $s_i \in \mathbb{S}$ und $i \in \{0, \dots, \|\mathbb{S}\|\}$ und es gilt wie oben $\sum_{i=1}^{\|\mathbb{S}\|} p_{s_i}^a = 1$. Im ersten Schritt wird die sogenannte Varianz des Durchschnitts Var bzw. die mittlere qua-

drastische Abweichung vom Mittelwert für eine Wahrscheinlichkeitsverteilung, gegeben durch \vec{p}^a , berechnet:

$$Var(\vec{p}^a) = \frac{1}{\|\mathbb{S}\|} \sum_{i=1}^{\|\mathbb{S}\|} (p_{s_i}^a - \mu)^2 \quad \text{mit: } \mu = \frac{1}{\|\mathbb{S}\|} \sum_{i=1}^{\|\mathbb{S}\|} p_{s_i}^a = \frac{1}{\|\mathbb{S}\|} \quad (4.16)$$

Damit ergibt sich eine Standardabweichung des Durchschnitts σ als die positive Wurzel der obigen Varianz:

$$\sigma(\vec{p}^a) = \sqrt{Var(\vec{p}^a)} \quad (4.17)$$

Für eine maximale Varianz gilt: $\vec{p}_{max} = (p_{s_1}, p_{s_2}, \dots, p_{s_{\|\mathbb{S}\|}})$ mit $p_{s_i} = 1$ für beliebiges aber festes i und sei $\forall j \neq i : p_{s_j} = 0$. Durch Normierung der Standardabweichung mit $\sigma_{max} = \sigma(\vec{p}_{max})$ ergibt sich hier die relative Standard-Unsicherheit zu:

$$\bar{u}_a(\vec{p}^a) = \frac{\sigma(\vec{p}^a)}{\sigma_{max}} \in [0, 1] \quad (4.18)$$

Die so definierte relative Standard-Unsicherheit bezieht sich dabei auf konkrete Messwerte. Bei Messwerten besteht die geringste Unsicherheit dann, wenn auch die Standardabweichung am geringsten ist, die Werte also alle gleich sind. Analog dazu besteht bei Messwerten die größte Unsicherheit, wenn sich die Werte größtmöglich unterscheiden, die Standardabweichung also am größten ist. Da es sich bei den Werten der Sekundärstruktur-Zuordnungen um Wahrscheinlichkeiten handelt, verhält sich die Unsicherheit hier hingegen genau umgekehrt zur Standardabweichung. Die Zuordnung ist dann am sichersten, wenn eine Struktur am wahrscheinlichsten ist ($p = 100\%$) und alle anderen Strukturen eine Wahrscheinlichkeit von 0% besitzen. Dies entspricht dem Fall der größten Standardabweichung. Am unsichersten ist die Zuordnung, wenn alle Strukturtypen gleich wahrscheinlich sind. Dies wiederum entspricht dem Fall der geringsten Standardabweichung. Deshalb definiert sich die Unsicherheit bei Sekundärstruktur-Zuordnungen schließlich wie folgt:

$$u_a^{Sec}(\vec{p}^a) = 1 - \bar{u}_a(\vec{p}^a) \quad (4.19)$$

Für eine Strukturzuordnung mit minimaler Unsicherheit gilt demnach: $u_a^{Sec}(\vec{p}_{max}) = 0$. Mit $\vec{p}_{min} = (p_{s_1}, p_{s_2}, \dots, p_{s_{\|\mathbb{S}\|}})$ mit $p_{s_i} = \frac{1}{\|\mathbb{S}\|}$ gilt für eine Strukturzuordnung mit maximaler Unsicherheit entsprechend: $u_a^{Sec}(\vec{p}_{min}) = 1$.

Kapitel 5

Visualisierung

Seeing is believing.

– *Englisches Sprichwort*

Seeing is not believing.

– *John Darius*

In diesem Kapitel werden die Visualisierungen für die Unsicherheit bei der Sekundärstruktur-Zuordnung beschrieben. Ausgangspunkt sind dabei die im vorangegangenen Kapitel 4 berechneten Unsicherheitswerte je Aminosäure und auch die Strukturtyp-Wahrscheinlichkeiten - sowohl auf der Ebene der Verfahren als auch auf Aminosäuren-Ebene. Im ersten Abschnitt erfolgt eine ausführliche Analyse vorhandener Darstellungskonzepte. Basierend auf dieser Auswertung erfolgte die Auswahl geeigneter Visualisierungen für den hier gegebenen Kontext. Anschließend wird dann die Umsetzung und die Implementierung der jeweils gewählten Darstellungen beschrieben. Entsprechend der Visualisierungspipeline in Abbildung 2.1 auf Seite 9 entsteht bei der Umsetzung der abgeleiteten Daten in eine Darstellung auch eine sogenannte Visualisierungsunsicherheit. Sie bringt zum Ausdruck, dass sowohl durch die Komplexität der Implementierung als auch durch die Art der Darstellung zusätzlich Unsicherheit mit einfließen kann. Diese Umstände sind allerdings vorher schwer abzuschätzen. Sie wurden insofern berücksichtigt, als dass für die Umsetzung der gewählten Darstellungen möglichst gut nachvollziehbare Algorithmen verwendet wurden. Eine Analyse dieser möglichen zusätzlichen Unsicherheit erfolgt im anschließenden Kapitel in der Auswertung.

5.1 Analyse geeigneter Darstellungskonzepte

In diesem Abschnitt wird die Entscheidung für die hier gewählten Darstellungen begründet. Für mögliche Visualisierungen von Unsicherheit stehen unzählige Methoden zur Verfügung

(siehe z.B. [PWL97]). Um die Wahl geeigneter Darstellungen zu unterstützen werden verschiedene Typologien bzw. Taxonomien vorgeschlagen (siehe dazu auch Abschnitt 3.3). Diese Klassifikationen stellen allerdings keine Bewertung der Visualisierungen dar. Bei nur wenigen Darstellungen lag bisher die Evaluation der praktischen Brauchbarkeit im Fokus, wie beispielsweise in [PRJ12], [GS06] und [ZC06] festgestellt wird. Die Entscheidung für eine bestimmte Unsicherheitsdarstellung ist stark kontextabhängig. Wie in Abschnitt 3.3 bereits ausgeführt, konnte keine Darstellung ermittelt werden, welche dem hier gegebene Kontext der Sekundärstruktur-Zuordnung bei Proteinen entspricht. Es kann hier somit nicht direkt auf eine vorhandene Darstellung zurückgegriffen beziehungsweise auf einer aufgebaut werden. Betrachtet man die Dimensionen der Daten, so handelt es sich bei den reduzierten Unsicherheitswerten um Skalare und bei den Strukturtyp-Wahrscheinlichkeiten um eindimensionale Vektoren. Ausgehend von der Dimension der darzustellenden Daten, werden, entsprechend der Taxonomie von [ZC06], Arbeiten mit evaluierten und hier anwendbaren Darstellungen berücksichtigt:

Anhand visueller Semiotik erfolgt bei [MRO+12] eine umfangreiche empirische Untersuchung punktförmiger Symbole, welche allgemeine Unsicherheit repräsentieren, in Bezug auf intuitive Wahrnehmung. Basierend auf den sieben von Bertin eingeführten visuellen Variablen werden hier für die Analyse zusätzlich vier weitere definiert: Position, Größe, Farbton, Helligkeit, Farbsättigung, Ausrichtung, Körnigkeit, Anordnung, Form, Transparenz und Unschärfe. Ausgehend von den Beschränkungen der durchgeführten Benutzerstudien werden in Bezug auf die Anwendbarkeit der Ergebnisse folgende hier relevante Aussagen gemacht: Unschärfe und Position eignen sich besonders gut. Helligkeit und Anordnung werden auch hoch bewertet. Sowohl Größe als auch Transparenz sind potentiell brauchbar. Farbsättigung wird hingegen recht niedrig bewertet.

Guo, Huang und Laidlaw [GHL15] untersuchen die Unsicherheit anhand verschiedener Repräsentationen für Kanten in Graphen, beziehen sich also auf eine linienbasierte Repräsentation von Unsicherheit. Dabei wurde die gegenseitige Beeinflussung verschiedener visueller Variablen analysiert. Die Effektivität einer visuellen Variablen wird daran gemessen, inwiefern sie sich zur Darstellung von Unsicherheit in Bezug auf die Wahrnehmung von Benutzern eignet. Für die Darstellung der Unsicherheit verwenden sie Helligkeit, Transparenz, Körnigkeit und Unschärfe. Verglichen wird mit der gleichzeitigen Belegung einer der visuellen Variablen Breite, Farbton oder Farbsättigung. Daraus resultierten folgende Entwurfsvorschläge für eine Unsicherheitsdarstellung: Helligkeit ist eine effektive visuelle Variable um Unsicherheit darzustellen. Allerdings wird davon abgeraten sie zusammen mit Variablen zu verwenden, welche durch den Farbton kodiert sind. Unschärfe, Körnigkeit und Transparenz eignen sich in der Verwendung als zweite Dimension hingegen gut. Unschärfe beeinflusst die Wahrnehmung von Breite negativ. Wenn zwei visuelle Variablen gleichzeitig für die Darstellung von Daten verwendet werden, kann die Wahrnehmung dadurch erleichtert werden, indem man die Unterscheidbarkeit der einen Variablen erhöht oder der andern erniedrigt. Unter der Unterscheidbarkeit wird dabei die wahrnehmbare Distanz zwischen zwei aufeinanderfolgenden Werten einer visuellen Variablen verstanden.

Leitner und Buttenfield [LB00] kodieren Unsicherheit anhand der visuellen Variablen Helligkeit, Textur und Sättigung. Ihre Untersuchung ergibt, dass Helligkeit zu den meisten richtigen Antworten führt. Textur bekommt weniger richtige Antworten und Sättigung die wenigsten.

Sanyal et al. [SZB+09] analysieren in einer Benutzerstudie für ein- und zweidimensionale Datensätze die vier üblicherweise verwendete Darstellungen für Unsicherheit: die Größe von Glyphen, die Farbe von Glyphen, die Farbe der Datenoberfläche und traditionelle Fehlerbalken. Die Fehlerbalken schneiden dabei deutlich schlechter ab als die Glyphen-Repräsentation. Als mögliche Gründe werden die Größe der jeweils in Anspruch genommenen Fläche der jeweiligen Repräsentation, sowie die feststehende Datendichte genannt. Unter den anderen drei Darstellungen wird keine konsistente Reihenfolge in Bezug auf Effektivität festgestellt. Es ergibt sich allerdings eine hohe Abhängigkeit der Effektivität von den unterschiedlichen Aufgaben, die den Benutzern gestellt werden. Die Effektivität ist also kontextabhängig.

Bei Boukhelifa et al. [BBIF12] wird für die Linienrepräsentation die *Skizzenhaftigkeit* als mögliche zusätzliche visuelle Variable untersucht. Bei der Evaluation durch Benutzer steht dabei die Ermittlung im Vordergrund, ob Skizzenhaftigkeit intuitiv mit Unsicherheit assoziiert wird. Verglichen wird mit den visuellen Variablen Unschärfe, Graustufen und Strichelung. Als bevorzugte Darstellung ergab sich Strichelung, gefolgt von Unschärfe, Graustufen und dann Skizzenhaftigkeit. Der Grund für die Bevorzugung von Strichelung liegt demnach in der guten Wahrnehmbarkeit, für Unschärfe liegt er in der Übereinstimmung damit, was Unsicherheit vermittelt und für die Graustufen liegt er in der einfachen Verständlichkeit. Die Semantik von Skizzenhaftigkeit liegt demnach nicht in der Unsicherheit, zumindest nicht ohne dass explizit auf die Bedeutung hingewiesen wird.

In [GR04] wird ein auf Punkten basierender Ansatz verwendet, um Oberflächen-Unsicherheit darzustellen. Dadurch bleiben die visuellen Variablen der Farbattribute unbelegt und stehen für andere Variablen zur Verfügung. Im Vergleich zu einer Pseudoeinfärbung der Oberflächen ergibt eine Benutzerstudie, dass der auf Punkten basierende Ansatz in Bezug auf Genauigkeit, Benutzerfreundlichkeit, allgemeine Zufriedenheit und Vertrauen in die Antwort besser abschneidet. Zu den Vorteilen zählen die Autoren auch, dass diese Art der Unsicherheitsdarstellung nicht ablenkt und nicht in die Oberflächeninformation eingreift. Allerdings ist zu berücksichtigen, dass sich die Unsicherheit hier auf die Position der Oberfläche bezieht.

In [ZC06] werden acht Darstellungen von Unsicherheit aus den unterschiedlichsten Bereichen anhand der theoretischen Grundsätzen zur Wahrnehmungstheorie von Bertin, Tufte und Ware analysiert. Von den verwendeten theoretischen Grundsätzen werden für die Auswertung Heuristiken abgeleitet. Für diese Heuristiken wird überprüft, in wie vielen Fällen sie für eine jeweilige Darstellung von Unsicherheit geeignet sind. Die Heuristik „*Stelle sicher, dass visuelle Variablen von ausreichender Länge sind*“ ist demnach in sieben von den acht untersuchten Fällen geeignet. Außerdem ist in sechs Fällen „*Binde Text ein, wann immer es relevant ist*“ dienlich. Noch in mindestens drei Darstellungen sind „*Entferne Belangloses*“, „*Präattentive Unterstützung nimmt mit der Größe des Blickfeldes zu*“ und „*Quantitative Abschätzung erfordert eine Änderung*“

der Position oder Größe“ von Bedeutung.¹ Nur noch in höchstens zwei Fällen dienlich sind die Heuristiken „Stelle mehrere Detailgrade zur Verfügung“, „Platziere die größtmögliche Menge an Daten auf kleinstmöglichem Raum“, „Erhalte die Daten innerhalb der grafischen Räumlichkeit“, „Erwarte keine Ablesereihenfolge von Farben“, „Farbwahrnehmung ändert sich mit der Größe von farbigen Objekten“, „Berücksichtige die Gestalt-Gesetze“, „Berücksichtige Farbenblindheit“ und „Lokaler Kontrast beeinflusst die Farb- und Graustufenwahrnehmung“.

Die visuellen Variablen Transparenz und Unschärfe werden häufig für die Darstellung von Unsicherheit in Bezug auf Aufenthaltswahrscheinlichkeit, Position oder Existenz verwendet. Da sich hier die Unsicherheit allerdings ausschließlich auf die Sekundärstruktur-Zuordnung bezieht, eigneten sich diese visuellen Variablen vor allem in der Cartoon-Darstellung nicht ohne Weiteres. Ein weiteres Ziel war es also auch, die Darstellungen so zu wählen, dass falsche Interpretationen und somit falsche Entscheidungen durch Benutzer vermieden werden. Eine Einschränkung bei der Wahl der Darstellungen ergab sich dadurch, dass die visuelle Variable Farbe in der Cartoon-Darstellung für die Repräsentation anderer Eigenschaften eines Proteins schon belegt ist. Außerdem gilt, je komplexer eine Unsicherheitsdarstellung, desto schwieriger ist sie zu interpretieren. Dies soll anhand des folgenden Zitats von Martin Krzywinski verdeutlicht werden: „Wenn etwas so einfaches wie Fehlerbalken schon missverstanden wird, so geschieht dies wahrscheinlich auch bei allem, was komplexer ist.“ Aus [Mar13] in Bezug auf [NS12].² Zudem handelt es sich bei Proteinen generell schon um komplexe dreidimensionale Strukturen. Im Sequenz-Diagramm und in der Cartoon-Darstellung muss davon ausgegangen werden, dass parallel zur Darstellung der Unsicherheit noch weitere Informationen eines Proteins dargestellt werden. Ein weiterer Aspekt bei der Konzentration auf eher einfache Darstellungen war deshalb auch, dass es für den Benutzer durch die zusätzliche Information der Unsicherheit nicht zu einer visuellen Überbeanspruchung kommt.

Die visuelle Wahrnehmung und Interpretation sind benutzerabhängig; zudem gibt es wie oben erwähnt bisher nur wenige Aussagen über die Eignung von verschiedenen Unsicherheitsdarstellungen für den hier gegebenen Kontext. Diesen Umständen wurde dadurch Rechnung getragen, dass nicht nur eine einzige Darstellung implementiert wurde. Für eine spätere Evaluation wurde außerdem mittels zahlreicher Parameter die Möglichkeit geschaffen, die einzelnen Darstellungen interaktiv auszuwählen, zu kombinieren und einzelne Eigenschaften der Darstellungen anzupassen.

¹Unter *präattentiver Wahrnehmung* versteht man die unterbewusste Wahrnehmung von Sinnesreizen, welche nicht ins Bewusstsein vordringen.

²Martin Krzywinski arbeitet am Canada's Michael Smith Genome Sciences Center in Vancouver, Kanada und beschäftigt sich dort mit Datenvisualisierung, beispielsweise in [Krz09].

5.2 Sequenz-Diagramm

Im zweidimensionalen Sequenz-Diagramm erfolgt die Betrachtung eines Proteins auf der Ebene der Aminosäuren-Sequenz, der Primärstruktur. Es dient dazu, verschiedene pro Aminosäure geltenden Eigenschaften gleichzeitig darzustellen. Die eindimensionale Sequenz bildet dabei die horizontale Achse. Parallel dazu werden im Diagramm zeilenweise die Eigenschaften aufgetragen. Die Implementierung des Sequenz-Diagramms erfolgt innerhalb des *Protein-Uncertainty-Plugins* im Modul *UncertaintySequenceRenderer* und basiert auf einer vorhandenen Implementierung, welche im Rahmen dieser Arbeit entsprechend angepasst und erweitert wurde. Für die visuelle Repräsentation einer Eigenschaft steht pro Aminosäure ein Feld von fester Breite zur Verfügung. Um aufgrund mehrerer gleichzeitig darstellbarer Eigenschaften die Übersichtlichkeit nicht einzuschränken, wurde die Höhe der einzelnen Felder auf das Maß der Breite begrenzt. Dadurch steht für die Visualisierung jeder Eigenschaft pro Aminosäure eine quadratische Kachel zu Verfügung. Der Einfachheit halber gilt für das Quadrat eine Kantenlänge von 1. Ausgehend von der zugrundeliegenden Implementierung einer Zeile für die Darstellung der Sekundärstruktur von STRIDE wurden noch Zeilen für PDB und DSSP hinzugefügt. Auch die Unsicherheitswerte (siehe Abschnitt 4.5), die Strukturtyp-Wahrscheinlichkeiten (siehe Abschnitt 4.4.5) und die Schwellenwert von STRIDE (siehe Abschnitt 4.3.2) werden jeweils in einer weiteren Zeile angezeigt. Die einzelnen Zeilen können auch ausgeblendet werden. In Weiteren, nicht ausblendbaren Zeilen, stehen die folgenden grundlegenden Eigenschaften eines Proteins zur Verfügung: PDB-Index, Ketten-Identifikator, Art der Aminosäure im Ein-Buchstaben-Code und Bindestellen von Enzymen. Bei fehlenden oder heterogenen Aminosäuren wird der Ein-

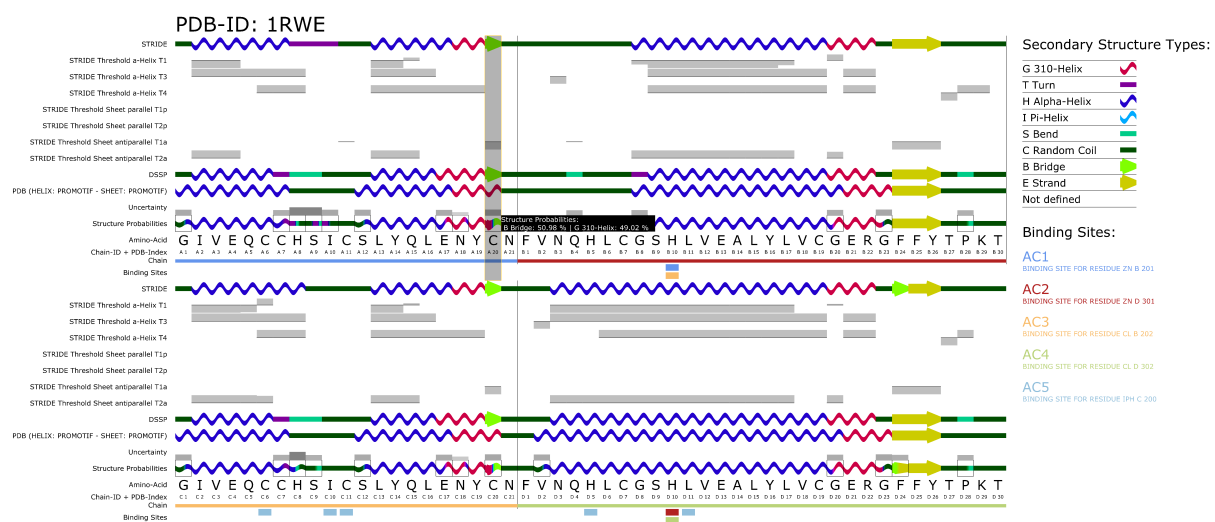


Abbildung 5.1: Vollständiges Sequenz-Diagramm für das Protein mit der PDB-ID 1RWE.

Buchstaben-Code einer Aminosäure farblich verändert. Diese für das Sequenz-Diagramm notwendigen Informationen sind (mit Ausnahme der Bindestellen, welche über ein separates Modul geladen werden können) alle in der Ausgabedatei der vorangegangenen Datenerfassung

enthalten. An der linken Seite des Diagramms findet sich die Beschriftung der Zeilen und am rechten Rand wird die Legende der Sekundärstrukturen angezeigt. Um die Sequenz der Ansicht am Bildschirm anpassen zu können, erfolgt nach einer gewünschten Anzahl an Aminosäuren ein Zeilenumbruch. Die Auswahl und Markierung bestimmter Aminosäuren ist ebenfalls möglich. Eine Ergänzung der visuellen Information um zusätzlichen kontextbezogenen Text erfolgt durch sogenannte *Tooltips*. Abhängig davon über welcher Eigenschaftszeile sich der Mauszeiger befindet, werden die tatsächlichen Werte der Eigenschaften einer Aminosäure eingeblendet.

5.2.1 Struktur-Morphing

Bei der geometrischen Repräsentation der Strukturtypen werden diese in drei Strukturformen S_{Form} zusammengefasst. Dabei werden die drei unterschiedlichen Helix-Typen (α , π und 3_{10}) durch eine gemeinsame Strukturform abgebildet, β -Strang und β -Brücken werden in einer zweiten Strukturform zusammengefasst und die verbleibenden Strukturtypen (Biegung, Umdrehung und zufällige Windung) ergeben eine dritte Strukturform. Die *Helix*-Strukturform wird dabei pro Kachel für einen Winkel $\alpha \in [0, 2\pi[$ über eine Sinus-Kurve $\sin(\alpha)$ dargestellt. Die *Strang*-Strukturform wird durch einen breiten Pfeil abgebildet, wobei jeweils nur die letzte Aminosäure eines β -Stranges durch die Pfeilspitze dargestellt wird. Die dritte Strukturform wird anhand einer horizontalen Linie repräsentiert. Die jeweilige Dicke h der Strukturformen

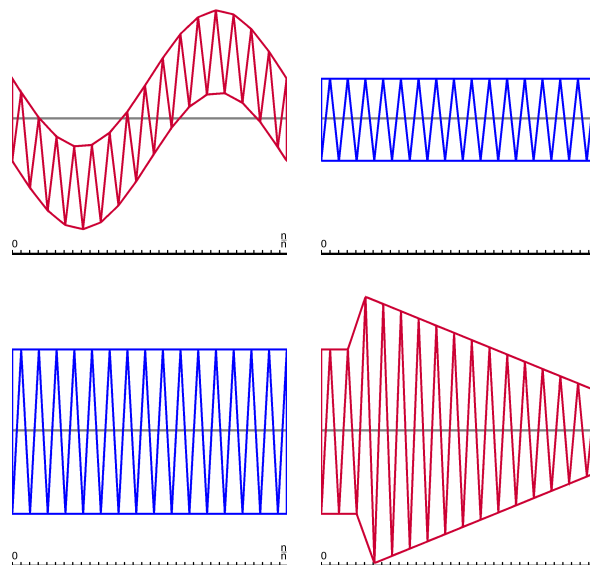


Abbildung 5.2: Unterteilung der unterschiedlichen Strukturformen in Dreiecksprimitiven. Von links oben nach rechts unten: Sinus-Kurve für Helices; schmaler Balken für Biegung, Umdrehung und zufällige Windung; breiter Balken für β -Strang; Pfeilspitze für β -Brücken und letzte Aminosäure eines β -Strangs.

wird nach der Sichtbarkeit in der Gesamtansicht des Sequenz-Diagramms festgelegt. Für einen Strukturtyp $s \in \mathbb{S}$ sei nachstehend die dazugehörige Strukturform durch $S_{Form}(s)$ gegeben. Anhand dieser Strukturformen werden die Sekundärstrukturen der einzelnen Verfahren in den jeweiligen Eigenschaftszeilen dargestellt.

Beim hier beschriebenen Struktur-Morphing werden für Aminosäuren mit unsicherer Strukturzuordnung die Strukturtyp-Wahrscheinlichkeiten anhand der visuellen Variablen *Form* ausgedrückt. Der Übergang von einer Strukturform zur anderen erfolgt proportional zur jeweiligen Strukturtyp-Wahrscheinlichkeit. Diese Darstellung wird in der für die Strukturtyp-Wahrscheinlichkeiten vorgesehenen Eigenschaftszeile im Sequenz-Diagramm abgebildet. Um die geometrischen Formen fließend ineinander übergehen lassen zu können, musste die in der ursprünglichen Implementierung verwendete Darstellung mittels Texturen durch eine mit geometrischen Primitiven ersetzt werden. Die Anzahl der verwendeten Dreiecksprimitiven n für die Strukturformen kann dabei aus Performanzgründen interaktiv verändert werden. Mittels der OpenGL-Primitive `TRIANGLE_STRIP` werden um die Mittellinie der quadratischen Kachel mit $\pm \frac{h}{2}$ alternierend und in Abständen von $\frac{1}{n}$ für jede Strukturform die Eckpunkte der Dreiecke $P(x, f(x)_{S_{Form}(s)})$ berechnet (siehe Abbildung 5.2). Dabei gilt für $x \in \mathbb{X} = \{0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n}\}$. Dadurch werden die zur Mittellinie symmetrischen Strukturformen geometrisch erzeugt. Die Eckpunkte werden beim Programmstart für jede Strukturform vorberechnet. Um für eine beliebige aber feste Aminosäure a unterschiedliche Strukturformen entsprechend der Strukturtyp-Wahrscheinlichkeiten $p_A(s_{[a]})$ ineinander übergehen zu lassen, müssen die Eckpunkte der Dreiecke verändert werden. Die resultierenden Eckpunkte seien dann gegeben durch $P(x, f(x)_{Morph})$. Dabei bleiben die Abstände in x -Richtung gleich und lediglich in y -Richtung ergeben sich die neue Werte $f(x)_{Morph}$. Mit $p_A(s_{[a]}) \in [0, 1]$ und $f(x)_{S_{Form}(s_{[a]})} \in [0, 1]$ gilt:

$$\forall x \in X : f(x)_{Morph} = \sum_{s \in \mathbb{S}} \left(p_A(s_{[a]}) \cdot f(x)_{S_{Form}(s_{[a]})} \right) \quad (5.1)$$

Für die Aminosäuren mit sicherer Strukturzuordnung werden die zutreffenden Strukturformen jeweils unverändert dargestellt. Als zusätzliche Option können diese allerdings ausgeblendet werden, wodurch nur noch die unsicheren Strukturzuordnungen anhand der ineinander übergehenden Strukturformen zu sehen sind.

5.2.2 Farbinterpolation im Balkendiagramm

Als weitere visuelle Variable für die Strukturtyp-Wahrscheinlichkeiten wurde der *Farbton* verwendet. Jeder Strukturtyp entspricht demnach einer fest zugewiesenen Strukturfarbe. Dies erlaubt im Gegensatz zur obigen Strukturform eine Differenzierung jedes einzelnen Strukturtyps. Verwendet wird diese Zuordnung auch zusammen mit den Strukturformen zur Darstellung der Sekundärstrukturen der einzelnen Verfahren in deren jeweiligen Eigenschaftszeilen.

In der Eigenschaftszeile für die Strukturtyp-Wahrscheinlichkeiten wurden diese mit den Strukturfarben qualitativ abgebildet. Dazu wurde jede Kachel einer Aminosäure entsprechend den Wahrscheinlichkeiten vertikal in Intervalle aufgeteilt, welche jeweils wiederum mit der dazugehörigen Strukturfarbe eingefärbt wurden. Dies entspricht einer breitenproportionalen Darstellungsform, einem sogenannten gestapelten Balkendiagramm, und gibt die Wahrscheinlichkeiten quantitativ wieder. Die Strukturformen werden anhand des so ermittelten Farbverlaufs eingefärbt. Wahlweise können die sicheren Zuordnungen ausgeblendet werden. Um durch die nahtlose Aneinanderreihung der Balkendiagramme die Unterscheidbarkeit der diskreten Aminosäuren aufrecht zu erhalten, können die Kacheln der unsicheren Zuordnungen mit einer Umrandung gegeneinander abgegrenzt werden.

Sortierung der Balken

Aufgrund wechselnder Strukturzuordnungen und bis zu vier Balken in verschiedenen Strukturfarben pro Aminosäure kommt es zu häufigen Farbwechseln. Um eine dadurch beeinträchtigte Interpretation zu verhindern, wurden die Balken der einzelnen Strukturtypen umgeordnet. Für eine beliebige aber feste Aminosäure a wurden alle Strukturtypen mit einer Zuordnungswahrscheinlichkeit größer als Null betrachtet: $s_{[a]} : p(s_{[a]}) > 0$. Für diese Strukturtypen wurde dann jeweils überprüft:

Besitzt die vorangehende Aminosäure $a - 1$ eine sichere Zuordnung $s_{[a-1]} : p(s_{[a-1]}) = 1$?

Wenn *JA*: Ist $s_{[a-1]} = s_{[a]}$?

Wenn *JA*: Sortiere $s_{[a]}^{p>0}$ an der ersten Position des Balkendiagramms ein.

Wenn *NEIN*: Ist $s_{[a]}$ = dem an letzter Position einsortierten Strukturtyp an $a - 1$?

Wenn *JA*: Sortiere $s_{[a]}$ an der ersten Position des Balkendiagramms ein.

Besitzt die folgende Aminosäure $a + 1$ eine sichere Zuordnung $s_{[a+1]} : p(s_{[a+1]}) = 1$?

Wenn *JA*: Ist $s_{[a+1]} = s_{[a]}$?

Wenn *JA*: Sortiere $s_{[a]}$ an der letzten Position des Balkendiagramms ein.

Ansonsten werden die verbleibenden Strukturtypen entsprechend absteigender Wahrscheinlichkeiten sortiert. Dadurch wird erreicht, dass gleiche Strukturtypen in aufeinanderfolgenden Zuordnungen an den Rändern der Balkendiagramme abgebildet werden. Somit wird die Anzahl der Farbübergänge deutlich reduziert.

Berücksichtigung unterschiedlicher Farbwahrnehmung

Die Zuordenbarkeit der Sekundärstrukturen anhand des Farbtons ist für die darauf basierende Interpretation und Entscheidungsfindung von fundamentaler Bedeutung. Deshalb lag bei der Bestimmung der Strukturfarben ein Fokus darauf, die Farben so zu wählen, dass auch von Farbenfehlsichtigkeit (z.B. Grünblindheit, Rotblindheit oder Blaublindheit) und Farbenblindheit

(Achromasie oder Monochromatie) betroffene Menschen die Strukturfarben gut unterscheiden können. Der Anteil davon betroffener Frauen liegt bei ca. 0.5%, der von Männern liegt mit ca. 8% deutlich höher [Cat00]. Begründet wird die Berücksichtigung zudem durch Wahrnehmungstheoretische Grundsätze (siehe dazu die Ausführungen zu [ZC06] in Abschnitt 5.1). Die Unterscheidbarkeit wird zu diesem Zweck nicht nur durch die Wahl des Farbtons, sondern auch zusätzliche durch die Variation der Helligkeit erhöht. Überprüft wurde die Farbzusammenhang durch eine Konvertierung der Farben in die Sichtweise verschiedener Farbenfehlsichtigkeiten. Auf der Webseite *VisCheck* kann z.B. eine solche Konvertierung vorgenommen werden.³









Sekundärstrukturen	HSL (0°-360°, 0-100%, 0-100%)	RGB (0-255, 0-255, 0-255)	Strukturfarben und -formen
π -Helix (5-Turn)	(340°, 100%, 40%)	(204, 0, 68)	
Umdrehung (Turn)	(290°, 100%, 30%)	(127, 0, 153)	
α -Helix (4-Turn)	(250°, 100%, 40%)	(34, 0, 204)	
3_{10} -Helix (3-Turn)	(200°, 100%, 50%)	(0, 170, 255)	
Biegung (Bend)	(160°, 100%, 40%)	(0, 204, 136)	
Zufällige Windung (Coil, Loop)	(120°, 100%, 15%)	(0, 77, 0)	
β -Brücke (Bridge)	(90°, 100%, 50%)	(128, 255, 0)	
(Erweiterter) β -Strang (Strand)	(60°, 100%, 40%)	(204, 204, 0)	

Tabelle 5.1: Die Strukturfarben in HSL- und RGB-Werten und in Kombination mit den Strukturformen.

Interpolation im HSL-Farbraum

Ein zusätzliches Kriterium bei der Farbwahl war, dass sich durch eine Interpolation der Strukturfarben indirekt die Unsicherheit einer Strukturzuordnung widerspiegelt. Dafür wurden die Farben im HSL-Farbraum betrachtet, da dieser im Vergleich zum RGB-Farbraum eine intuitivere und wahrnehmungsbezogenere Repräsentation darstellt [JG78]. Angegeben werden zur Definition einer Farbe im HSL-Farbraum der Farbton (*Hue*), die Sättigung (*Saturation*) und die Helligkeit (*Lightness*) in zylindrischen Koordinaten. Der Farbton wird als Winkel auf dem Farbkreis, die Sättigung und die Helligkeit werden prozentual angegeben. Die Funktionen für die Konvertierung der Farben zwischen den Farbräumen RGB und HSL basieren

³VisCheck: <http://www.vischeck.com>

auf von der Webseite *EasyRGB* entnommenen Algorithmen.⁴ Die Farbtöne wurden entsprechend der verfahrensunabhängigen Diskrepanz (siehe Abschnitt 4.4.1) im Farbkreis angeordnet. Bei der Interpolation zweier Strukturfarben im Farbkreis ergeben sich so - proportional zur Diskrepanz - unterschiedlich viele Zwischenfarbwerte. Würden die Farbwerte zwischen den

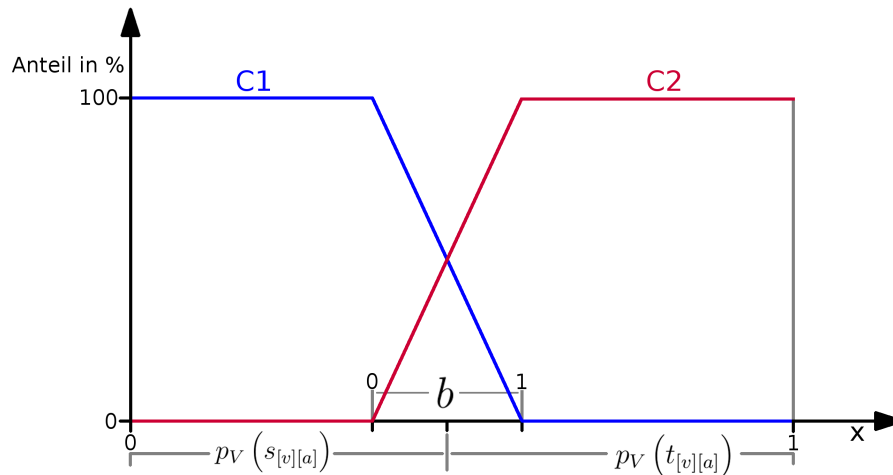


Abbildung 5.3: Interpolation zweier Strukturfarben C_1 und C_2 , je für die Strukturtyp-Wahrscheinlichkeiten $p_V(s_{[v][a]})$ und $p_V(t_{[v][a]})$ innerhalb des Intervalls b .

Intervallen gleichmäßig ineinander übergehen, wäre der proportionale Zusammenhang zu den Wahrscheinlichkeiten nicht mehr gegeben. Deshalb wird für die Interpolation ein beliebiges aber festes Intervalls der Breite b im Grenzbereich zwischen zwei Balken einer Kachel definiert, auf das sich die Interpolation beschränkt (siehe Abbildung 5.3). Die Breite des Intervalls wird auf den Wertebereich $[0, 1]$ normiert. Die Implementierung der Farbinterpolation erfolgt im Fragment-Shader. Für jede Kachel werden dem Fragment-Shader die Sortierung der Strukturtypen und die zugehörigen Wahrscheinlichkeiten übermittelt. Mittels der x -Koordinate wird nun wiederum für jedes Fragment überprüft, in welchem Balken es sich befindet und ihm entsprechend die Strukturfarbe zugewiesen. Befindet sich das Fragment allerdings im Grenzbereich zwischen zwei Balken innerhalb eines Interpolationsintervalls, ergibt sich die Farbe des Fragments anhand der beiden beteiligten Farben C_1 und C_2 entsprechend der normierten Breite. Die Helligkeit L für ein Fragment $Frage$ berechnet sich dann über: $L_{Frage} = b \cdot L_{C_1} + (1-b) \cdot L_{C_2}$. Die Sättigung S ergibt sich analog zu: $S_{Frage} = b \cdot S_{C_1} + (1-b) \cdot S_{C_2}$. Um die diskrepanzabhängigen Farbübergänge zu erhalten, darf die Interpolation des Farbtons H nur innerhalb einer festen Periode von $[0, 2\pi]$ erfolgen. Dies wird wie folgt erreicht: Wenn $H_{C_1} > H_{C_2}$ dann ist $H_{Frage} = H_{C_2} + b \cdot (H_{C_1} - H_{C_2})$. Für $H_{C_2} > H_{C_1}$ gilt dann analog $H_{Frage} = H_{C_1} + b \cdot (H_{C_2} - H_{C_1})$.

⁴EasyRGB: <http://easyrgb.com/index.php?X=MATH>

5.2.3 Säulendiagramm für Unsicherheits- und Schwellenwerte

Die Darstellung der reduzierten Unsicherheitswerte und der Schwellenwerte von STRIDE erfolgt nach dem Schema eines Säulendiagramms Abbildung 5.1. Bei den Unsicherheitswerten wird ausgehend vom unteren Rand einer Kachel der Wert in Richtung der positiven Ordinate nachse aufgetragen. Bei den Eigenschaftszeilen für die Schwellenwerte von STRIDE werden die jeweils berechneten Werte innerhalb der in Abschnitt 4.3.2 definierten Intervalle dargestellt. Die horizontale Mittellinie einer Kachel entspricht dabei dem jeweiligen Schwellenwert. Der Wertebereich einer Kachel ergibt sich über die in Tabelle 4.2 auf Seite 32 berechneten Intervallgrenzen. In Kacheln für welche STRIDE keine Werte berechnet hat, wird nichts angezeigt. Außer der visuellen Variablen *Länge*, welche durch die Höhe der Säulen gegeben ist, wird noch die visuelle Variable *Helligkeit* ausschließlich anhand unterschiedlicher Graustufen abgebildet (H ist beliebig, $S = 0$). Auf die Verwendung zusätzlicher Farben wurde verzichtet, da diese visuelle Variable durch die vorangehend genannte Darstellung schon belegt ist. Es gilt, je größer die Unsicherheit für eine Aminosäure an Position a , desto geringer ist die Helligkeit L der dargestellten Säule: $u_a^{Sec} \sim \frac{1}{L}$. Bei den Werten $t_{[v][a]}^i$, welche innerhalb eines Schwellenwertintervalls liegen, ist die Größe der Abweichung vom Schwellenwert dagegen proportional zur Helligkeit: $|t_{[v][a]}^i - T_i| \sim L$.

5.3 Cartoon-Darstellung

In diesem Abschnitt werden die Visualisierungen für die dreidimensionale Cartoon-Darstellung, einem sogenannten Bändermodell, beschrieben. Mit der Cartoon-Darstellung werden die Sekundärstrukturen eines Proteins als abstrakte Repräsentation einer Aminosäure abgebildet. Entsprechend der Gruppierung der Strukturtypen in drei Strukturformen wie in Abschnitt 5.2.1 werden in der dreidimensionalen Erweiterung Helices über breite Bänder dargestellt, wobei sich die helikale Krümmung aufgrund der dreidimensionalen Struktur des Proteins ergibt. Auch β -Stränge werden über Bänder repräsentiert, wobei die jeweils letzte Aminosäure davon genauso wie die β -Brücken über eine Pfeilspitze dargestellt werden. Alle anderen Strukturtypen werden über eine Röhre abgebildet. Die Position einer Sekundärstruktur wird durch die Position des C_α -Atoms der zugehörigen Aminosäure festgelegt. Zur detaillierten Darstellung der Strukturtypen muss in der Cartoon-Darstellung somit auch auf die in Abschnitt 5.2.2 definierten Strukturfarben zurückgegriffen werden. Allerdings kann die visuelle Variable, wie in der Aufgabenbeschreibung erwähnt, auch andere Proteineigenschaften abbilden. Auch können gleichzeitig zur Darstellung der Sekundärstruktur eines Proteins noch andere dreidimensionale Darstellungen überlagernd angezeigt werden, z.B. das Kalottenmodell (einzelne Atome werden durch Kugeln repräsentiert), das Gitter- oder Stäbchenmodell (Atombindungen werden in Stabform dargestellt) oder eine Oberflächendarstellung (z.B. die von einem Lösungsmittel erreichbare Oberfläche eines Moleküls). Die Unsicherheitsvisualisierungen für

die Cartoon-Darstellung wurden innerhalb des *Protein-Uncertainty-Plugins* im Modul *UncertaintyCartoonRenderer* implementiert. Die zugrunde liegende Cartoon-Darstellung, welche für die Aufgabe im Rahmen dieser Arbeit erweitert und angepasst wurde, basiert auf der Arbeit von Krone, Bidmon und Ertl [KBE08]. Die konkrete Implementierung erfolgte anhand einer modifizierten Variante des Verfahrens nach Hermosilla et al. [HG15], welche für die Berechnung der Darstellung *Tessellation-Shader* verwendet. Als lokales Beleuchtungsmodell wird das Verfahren nach Blinn-Phong verwendet. Die Darstellung dynamischer Positionsdaten eines Proteins wurde erhalten. Die vorhandene Implementierung lädt die dynamischen Proteindaten aus PDB-Dateien und zeigt dabei allerdings keine heterogenen oder fehlenden Aminosäuren an. Deshalb muss die Aminosäuresequenz der Unsicherheitsdaten, gegeben in der Eingabedatei der Datenerfassung, mit der tatsächlich angezeigten Sequenz für jeden *Render*-Aufruf anhand des PDB-Index synchronisiert werden.

5.3.1 Geometrieverzerrung

Über die Verzerrung der Oberfläche der Sekundärstruktur in der dreidimensionalen Darstellung werden, im Unterschied zu den Strukturtyp-Wahrscheinlichkeiten des Struktur-Morphings in Abschnitt 5.2.1, hier die Unsicherheitswerte über die visuelle Variable *Form* ausgedrückt. Die Berechnung der Oberflächengeometrie erfolgt dabei während der *Tessellation-Evaluation*-Stufe. Bei der Cartoon-Darstellung werden, um die Krümmung der Sekundärstruktur nachzubilden, jeweils die Positionen von vier aufeinander folgenden C_α -Atomen mittels kubischer B(asis)-Spline-Interpolation zu einer glatten Kurve verbunden. Die Ausrichtung der Kurve wird dabei durch die Richtung vom C_α -Atom zum O -Atom der Hauptkette bestimmt. Damit wird dann für jede Aminosäure das Dreiecksgitter der jeweiligen Strukturform erzeugt. Aus Gründen der Performanz kann der Unterteilungsgrad des Dreiecksgitters dabei interaktiv angepasst werden. Analog zu Textur-Koordinaten werden die erzeugten Eckpunkte der Dreiecke in der Tessellation-Evaluation-Stufe in der Ebene liegend in UV-Koordinaten angegeben. Die Positionen der Eckpunkte werden dann entsprechend der Strukturformen auf dreidimensionale Koordinaten abgebildet. Gegeben seien somit die berechneten 3D-Koordinaten der Eckpunkte \vec{e} und für die spätere Beleuchtungsberechnung deren senkrecht zur Oberfläche stehenden Normalenvektoren \vec{n} . Zudem sei die an dem Eckpunkt geltende Tangente mit \vec{t} gegeben. Ausgehend davon setzt die Geometrieverzerrung an. Da eine lineare Abbildung der Unsicherheitswerte gegeben sein sollte, wurden hier für die Verzerrung der Oberfläche periodische Funktionen gewählt. Diese sind aufgrund der linearen Skalierbarkeit der Amplitude und der Frequenz für diesen Zweck geeignet. Implementiert wurden die Sinus-Funktion $f(x)_{Sin}$ und eine periodisierte Dreiecksfunktion $f(x)_{Tri}$. Um den Einfluss dieser Darstellung auf die Gesamtansicht steuern zu können, sind für die Amplitude und Frequenz über Parameter noch zusätzliche feste Faktoren k_{amp} und k_{frq} wählbar. Die Funktionen können dabei ausgehend von der Betrachtung in UV-Koordinaten, auch zweidimensional übereinander gelagert dargestellt werden. Für die Koordinaten eines Eckpunktes u und v und im Folgenden mit $x \in \{u, v\}$, definieren sich

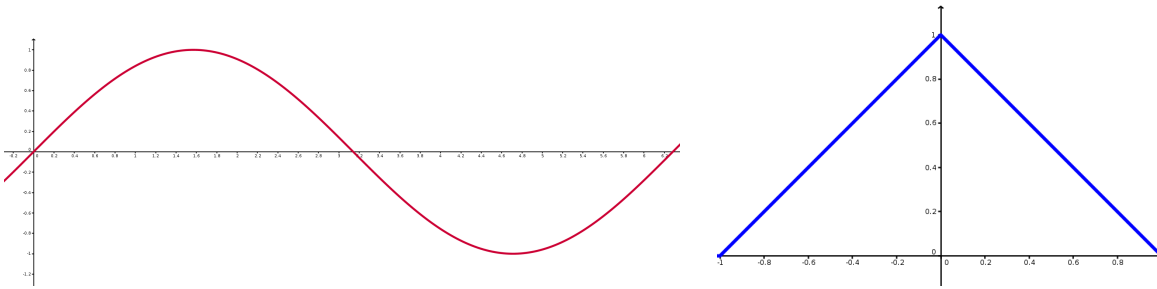


Abbildung 5.4: Links: Sinusfunktion (periodisch in $[0, 2\pi]$). Rechts: Dreiecksfunktion (periodisch in $[-1, 1]$).

die Funktionswerte für die beiden hier gewählten Funktionen für eine beliebige aber feste Aminosäure a und für Unsicherheitswerte $u_{[a]}$ jeweils zu:

$$f(x)_{Sin} = u_{[a]} \cdot k_{amp} \cdot \sin(2\pi \cdot x \cdot k_{frq} \cdot u_{[a]}) \quad (5.2)$$

$$f(x)_{Tri} = u_{[a]} \cdot k_{amp} \cdot \left(0.5 - \left|0.5 - \left(x \cdot k_{frq} \cdot u_{[a]} - \lfloor x \cdot k_{frq} \cdot u_{[a]} \rfloor\right)\right|\right) \quad (5.3)$$

Um den linearen Zusammenhang beizubehalten, werden die Funktionen für die Überlagerung in u - und v -Richtung mittels Addition gefaltet und ergeben die Funktion $g(u, v)_{Sin} = f(u)_{Sin} + f(v)_{Sin}$ bzw. $g(u, v)_{Tri} = f(u)_{Tri} + f(v)_{Tri}$. Die Verzerrung eines Eckpunktes \vec{e} erfolgt in Richtung seines Normalenvektors und berechnet sich demnach jeweils wie folgt:

$$\text{Für Sinusfunktion: } \vec{e}_{vzt} = \vec{e} + \vec{n} \cdot \left(g(u, v)_{Sin} + u_{[a]} \cdot k_{amp}\right) \quad (5.4)$$

$$\text{Für Dreiecksfunktion: } \vec{e}_{vzt} = \vec{e} + \vec{n} \cdot g(u, v)_{Tri} \quad (5.5)$$

Bei der Sinus-Funktion wird, um die ursprünglichen Dicke einer Strukturform und um mögliche invertierte Oberflächen zu vermeiden, der Wert der maximalen negativen Amplitude $u_{[a]} \cdot k_{amp}$ als Ausgleich dazu addiert.

Anschließend müssen noch die Normalenvektoren der verzerrten Oberfläche entsprechend angepasst werden. Der Normalenvektor für einen beliebigen aber festen Punkt steht dabei für eine zweidimensionale Funktion senkrecht zu der an diesem Punkt geltenden Tangentenebene. Die Tangentenebene wird dabei über die an dieser Stelle geltende Steigung der Funktion definiert. Nach der hier anwendbaren Summenregel der Differenzialrechnung gilt für die Steigung bzw. Ableitung einer Funktion $g(x, y)$:

$$\begin{aligned} g(x, y) &= f(x) + h(y) \\ g'(x, y) &= f'(x) + h'(y) \end{aligned}$$

Zusammen mit der Faktor-Regel $(k \cdot g(x))' = k \cdot g(x)'$ können die Ableitungen der Funktionen $f(x)_{Sin}$ und $f(x)_{Tri}$ (mit $x \in \{u, v\}$) für die Koordinaten u, v getrennt voneinander berechnet werden. Die in der zweidimensionalen Betrachtung berechneten Steigungen müssen anschließend in die dreidimensionale Betrachtung übertragen werden. Dafür wird noch senkrecht zur

$|\vec{h}| = 1$ folgt: $\alpha = \arccos(\vec{y} \cdot \vec{h})$. Mit der Winkelsumme ergibt sich dann $\beta = 180^\circ - 90^\circ - \alpha$. Durch die Anwendung des Sinussatz erhält man dann Δt :

$$\Delta t = \frac{|\vec{n}|}{\sin(\beta)} \cdot \sin(\alpha)$$

Für v erhält man analog dazu Δc . Damit setzt sich die resultierende normierte Normale \vec{n}_{vzt} für einen verzerrten Eckpunkt wie folgt zusammen:

$$\vec{n}_{vzt} = \frac{\vec{n} + \Delta t \cdot \vec{t} + \Delta c \cdot \vec{c}}{\|\vec{n} + \Delta t \cdot \vec{t} + \Delta c \cdot \vec{c}\|} \quad (5.6)$$

5.3.2 Konturen

Analog zur Geometrieverzerrung im letzten Abschnitt, werden über die Darstellung von Konturen die Unsicherheitswerte hier über die visuelle Variable *Breite* abgebildet. Auch hier ist pro Aminosäure ein linearer Zusammenhang zwischen der Unsicherheit und der Konturenbreite gegeben. Die Kontur wird schwarz dargestellt, wodurch ein Konflikt mit der visuellen Variablen *Farbe* vermieden wird. Als geeignetes Verfahren wird die beim *Cel-Shading* üblicherweise verwendete Vorgehensweise gewählt.⁵ Dabei wird von den ansonsten nicht sichtbaren und verworfenen Dreiecke der Rückseite ausgegangen. Es wird entweder die rückseitige Geometrie im Linien-Modus gezeichnet, wobei man über die Linienbreite auch die Breite der Kontur bestimmen kann. In der zweiten Variante wird die Geometrie ausgefüllt gezeichnet. Über eine Skalierung der Dreieckseckpunkte entlang der Normalenvektoren wird so eine Kontur erzeugt. Wie sich nach einer anfänglichen Implementierung herausstellte, ist die Darstellung der Konturen im Linien-Modus mit einem drastischen Performanzeinbruch verbunden. Deshalb wurde alternativ dazu übergegangen, die Dreiecke im ausfüllenden Modus zu zeichnen. Die vollständige Geometrie muss dabei ein zweites Mal mit invertiertem *Backface-Culling* gezeichnet werden. In OpenGL kann über die Funktion `glCullFace()` mit `GL_FRONT` oder `GL_BACK` angegeben werden, welche Dreiecke mit entsprechender Orientierung gezeichnet werden sollen. Da hier wieder die Position der Eckpunkte verändert wird, findet die Berechnung in der *Tessellation-Evaluation*-Stufe statt. Die Verschiebung der Eckpunkte \vec{e} in Richtung des dazugehörigen Normalenvektors \vec{n} berechnet sich dabei wie folgt:

$$\vec{e}_{knt} = \vec{e} + \vec{n} \cdot (k_{knt} \cdot u_{[a]}) \quad (5.7)$$

Zur Steuerung des Einflusses der Kontur auf die Gesamtdarstellung kann die Konturenbreite zusätzlich über einen globalen Skalierungsfaktor k_{knt} angepasst werden.

⁵Siehe beispielsweise: http://raulreyesfinalproject.files.wordpress.com/2012/12/dissertation_cell-shading-raul_reyes_luque.pdf

5.3.3 Screen-Door-Transparency

Zur Repräsentation der Strukturtyp-Wahrscheinlichkeiten in der dreidimensionalen Darstellung wird ein auf Punkten basierender Ansatz gewählt, welcher die visuellen Variablen *Transparenz* und *Körnigkeit* vereint. Zur Differenzierbarkeit der einzelnen Strukturtypen ist zusätzlich die Belegung der visuellen Variablen *Farbe* entsprechend der in Abschnitt 5.2.2 definierten Strukturfarben notwendig. Ohne die Verwendung der Strukturfarben ist nur eine eingeschränkte, rein geometrische Unterscheidung anhand der Strukturformen möglich. Die zugrundeliegende Vorgehensweise wird mit dem Begriff *Screen-Door-Transparency* beschrieben [FDFH95]. Konkret verwendet wird dafür die Methode des *Ordered Dithering*, ein sogenanntes Halbtonverfahren.⁶ Neben einer Vielzahl weiterer möglicher Dithering-Methoden eignet sich diese hier aufgrund ihrer interaktiven Umsetzbarkeit im Fragment-Shader. Mit einem solchen Verfahren kann durch Fehlerdiffusion mit einer verringerten Anzahl an zur Verfügung stehenden Farben der Eindruck einer größeren Farbtiefe erzeugt werden. Außerdem kann man diesen Effekt für die Darstellung übereinanderliegender transparenter Objekte verwenden. Im Gegensatz zum *Alpha-Blending* müssen die Objekte hier nicht aufwendig sortiert werden. Dabei findet die Kombination der Objektfarben nicht pro Pixel (Bildpunkt) statt. Anhand eines Kriteriums wird entschieden, welche Objektfarbe ein Pixel hat und aufgrund einer bestimmten resultierenden Abfolge der Objektfarben wird nur der Eindruck einer Farbmischung erzeugt. Diese diskrete Unterscheidung der Objektfarben führt zu einer körnigen Darstellung. Dieser Effekt ist hier allerdings gewollt, da die Körnigkeit proportional zur Unsicherheit einer Struktur-Zuordnung ist. Außerdem bezieht sich die transparente Darstellung nur auf die Strukturen einer Aminosäure und ist unabhängig von der darunterliegenden Geometrie, überdeckt sie somit. Dies war auch ein Kriterium bei der Wahl dieser Darstellung. Denn aufgrund der möglichen Komplexität von Proteinen wird bei transparenter Visualisierung die Anzahl der übereinanderliegenden Geometrie, welche noch unterscheidbar ist, häufig überschritten.

Im Weiteren wird die Implementierung des hier gewählten Verfahrens näher beschrieben. Die für eine Aminosäure in der entsprechenden Strukturfarbe gezeichnete dreidimensionale Strukturform erzeugt im Bildraum, abhängig von ihrer Orientierung und Lage im Raum, eine entsprechende Anzahl an Fragmenten. Die Anzahl der davon tatsächlich gezeichneten Fragmente entspricht der Zuordnungswahrscheinlichkeit des betrachteten Strukturtyps. Die Entscheidung, wann ein Fragment gezeichnet oder verworfen wird, hängt dabei von seiner Position im Bildraum ab. Um diese Entscheidung treffen zu können, werden die Strukturtyp-Wahrscheinlichkeiten einer Aminosäure a der Größe nach sortiert:

$$\forall s \in \mathbb{S} \text{ und } \forall 1 \leq i \leq \|\mathbb{S}\| : s_{[a]}^i \leq s_{[a]}^{i+1}$$

⁶Beschreibung verschiedener Dithering-Algorithmen: <http://www.efg2.com/Lab/Library/ImageProcessing/DHALF.TXT>

Anschließend wird für einen beliebigen aber festen Strukturtyp $s_{[a]}^i$ anhand seiner Strukturtyp-Wahrscheinlichkeit $p_A(s_{[a]}^i)$ ein Intervall $I = [I_{min}, I_{max}]$ mit folgenden Grenzen festgelegt:

$$I_{max} = 1 - \sum_{\forall j < i} p_A(s_{[a]}^j) \quad (5.8)$$

$$I_{min} = I_{max} - p_A(s_{[a]}^i) \quad (5.9)$$

Beim Strukturtyp mit der größten Wahrscheinlichkeit ergibt sich beispielsweise das Intervall: $[p_A(s_{[a]}^1), 1]$. Die Intervall-Grenzen für den aktuell gezeichneten Strukturtyp werden dem Fragment-Shader übergeben. Entsprechend dem *Ordered Dithering* wird für ein Fragment mit den Koordinaten x_{frag} und y_{frag} innerhalb einer nach Bayer [Bay99] definierten sogenannten Dither- oder Bayer-Matrix geprüft, ob ein bestimmter dort angegebener Wert innerhalb der Intervall-Grenzen des Strukturtyps liegt. Die Bayer-Matrix wird dabei als sich im Bildraum entsprechend der Pixelzahl in horizontaler und vertikaler Richtung wiederholend angenommen. Um eine ausreichende Genauigkeit zu gewährleisten und eine ungewünschte starke Musterbildung aufgrund der Wiederholung der Matrix im Bildraum zu vermeiden, wird hier eine 8×8 -Matrix verwendet. Die Werte einer Bayer-Matrix liegen im Intervall $[0, 1]$:

$$M_{Bayer8x8} = \frac{1}{64} \begin{bmatrix} 0 & 32 & 8 & 40 & 2 & 34 & 10 & 42 \\ 48 & 16 & 56 & 24 & 50 & 18 & 58 & 26 \\ 12 & 44 & 4 & 36 & 14 & 46 & 6 & 38 \\ 60 & 28 & 52 & 20 & 62 & 30 & 54 & 22 \\ 3 & 35 & 11 & 43 & 1 & 33 & 9 & 41 \\ 51 & 19 & 59 & 27 & 49 & 17 & 57 & 25 \\ 15 & 47 & 7 & 39 & 13 & 45 & 5 & 37 \\ 63 & 31 & 55 & 23 & 61 & 29 & 53 & 21 \end{bmatrix}$$

Der mit dem Strukturtyp-Intervall zu vergleichende Wert der Bayer-Matrix ergibt sich dann zu: $m_{x,y} = M_{Bayer8x8}[x_{frag} \bmod 8][y_{frag} \bmod 8]$. Ist die Bedingung $I_{min} \leq m_{x,y} \leq I_{max}$ erfüllt, wird das Fragment gezeichnet, ansonsten nicht. Mit den Intervallen I wird somit jedem Pixel immer nur die Strukturfarbe genau eines Strukturtyps zugeordnet.

Beginnend beim wahrscheinlichsten Strukturtyp muss die gesamte Proteinstruktur für alle weniger wahrscheinlichen jeweils erneut gezeichnet werden. Per Parameter ist einstellbar, wie viele Strukturtypen so übereinander gelegt gezeichnet werden sollen. Um den Aufwand der zusätzlichen Geometrie-Pässe zu verringern, wird so bald wie möglich die Berechnung abgebrochen, da alle sicheren und die jeweils wahrscheinlicheren Strukturen schon gezeichnet wurden und die Anzahl der zu zeichnenden Strukturen entsprechend abnimmt.

Kapitel 6

Ergebnisse

Im ersten Teil dieses Kapitel werden anhand ausgewählter Aminosäuren die Ergebnisse der Visualisierungen im Sequenz-Diagramm und in der Cartoon-Darstellung präsentiert. Anschließend wird die Performanz der Implementierung untersucht. In der folgenden Diskussion der Ergebnisse wird das Unsicherheitsmodell bewertet und die Visualisierungen werden evaluiert. Zusätzlich wird jeweils noch auf mögliche Erweiterungen eingegangen.

6.1 Visualisierungen der Unsicherheit

Für die Untersuchung der implementierten Darstellungen der Unsicherheit werden in den folgenden Abschnitten verschiedene Ansichten aufgeführt. Bei Proteinen handelt es sich um komplexe dreidimensionale Strukturen, wobei die durchschnittliche Größe eines Proteins ca. 300 Aminosäuren beträgt. Um die Darstellungen im Detail untersuchen zu können, wurde das relativ kleine Protein mit der PDB-ID 1RWE (menschliches Hormon, welches die Aktivität von Insulin verstärkt) und einer Anzahl von 102 Aminosäuren gewählt. Trotz seiner geringen Größe bietet dieses Protein eine ausreichend ausgeprägte Sekundärstruktur für eine differenzierte Betrachtung. Um die Sichtbarkeit und Interpretation der Unsicherheit auch bei größeren Proteinen untersuchen zu können, wurde zum einen das Protein 1TII (Enterotoxin bei *Escherichia coli*) mit einer durchschnittlichen Größe von 738 Aminosäuren und zum anderen das relativ große Protein 4DOM (menschliches Signal-Protein) mit 9996 Aminosäuren gewählt. Die Parameter der einzelnen Darstellungen wurden für die folgenden Ansichten abhängig vom Gesamteindruck gewählt. Die Breite des Intervalls für die Interpolation der Strukturfarben im Sequenz-Diagramm wurde auf 0, 2 festgelegt. Der Faktor für die Konturen liegt im Bereich 7-10. Der Faktor für die Frequenz der Geometrieverzerrung ist 1 und der Faktor für die Amplitude liegt im Bereich 1-3. Die Unterteilungsstufe der Geometrie liegt in beiden Ansichten bei 64.

6.1.1 Sequenz-Diagramm

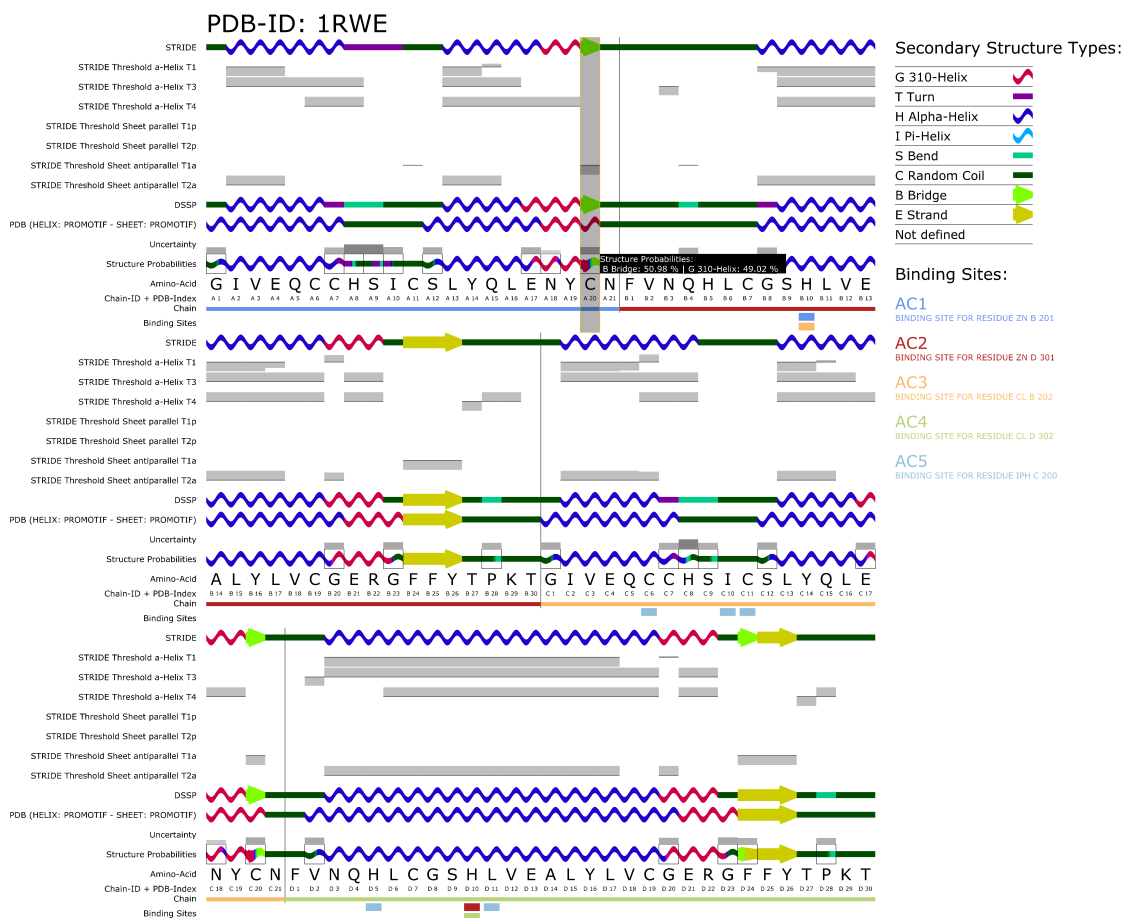


Abbildung 6.1: 1RWE: Vollständiges Sequenz-Diagramm.

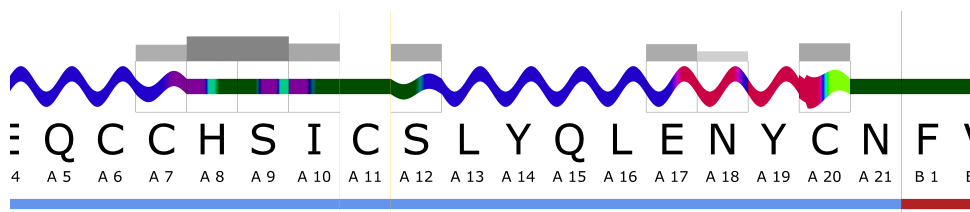


Abbildung 6.2: 1RWE: Vergrößerter Ausschnitt aus dem Sequenz-Diagramm. In den Farbverläufen tauchen zusätzliche Strukturfarben auf. Dies bringt die Diskrepanz der Strukturtypen zum Ausdruck.

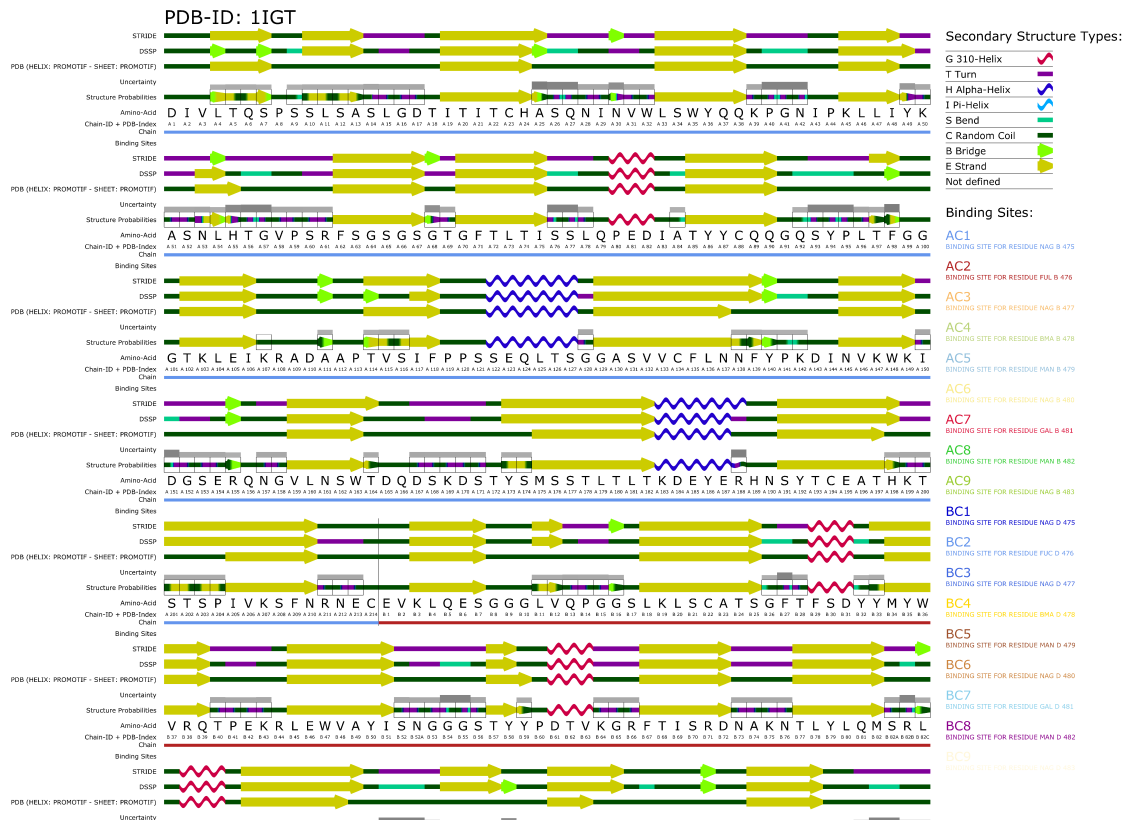


Abbildung 6.3: IIGT: Ausschnitt aus dem Sequenz-Diagramm, die Schwellenwerte von STRIDE sind ausgeblendet.

6.1.2 Cartoon-Darstellung



(a) Geometrieverzerrung in UV-Richtung mit Dreiecksfunktion. (b) Konturen mit unterschiedlicher Einfärbung der Aminosäuren.

Abbildung 6.4: 1RWE: Cartoon-Darstellungen. (1)

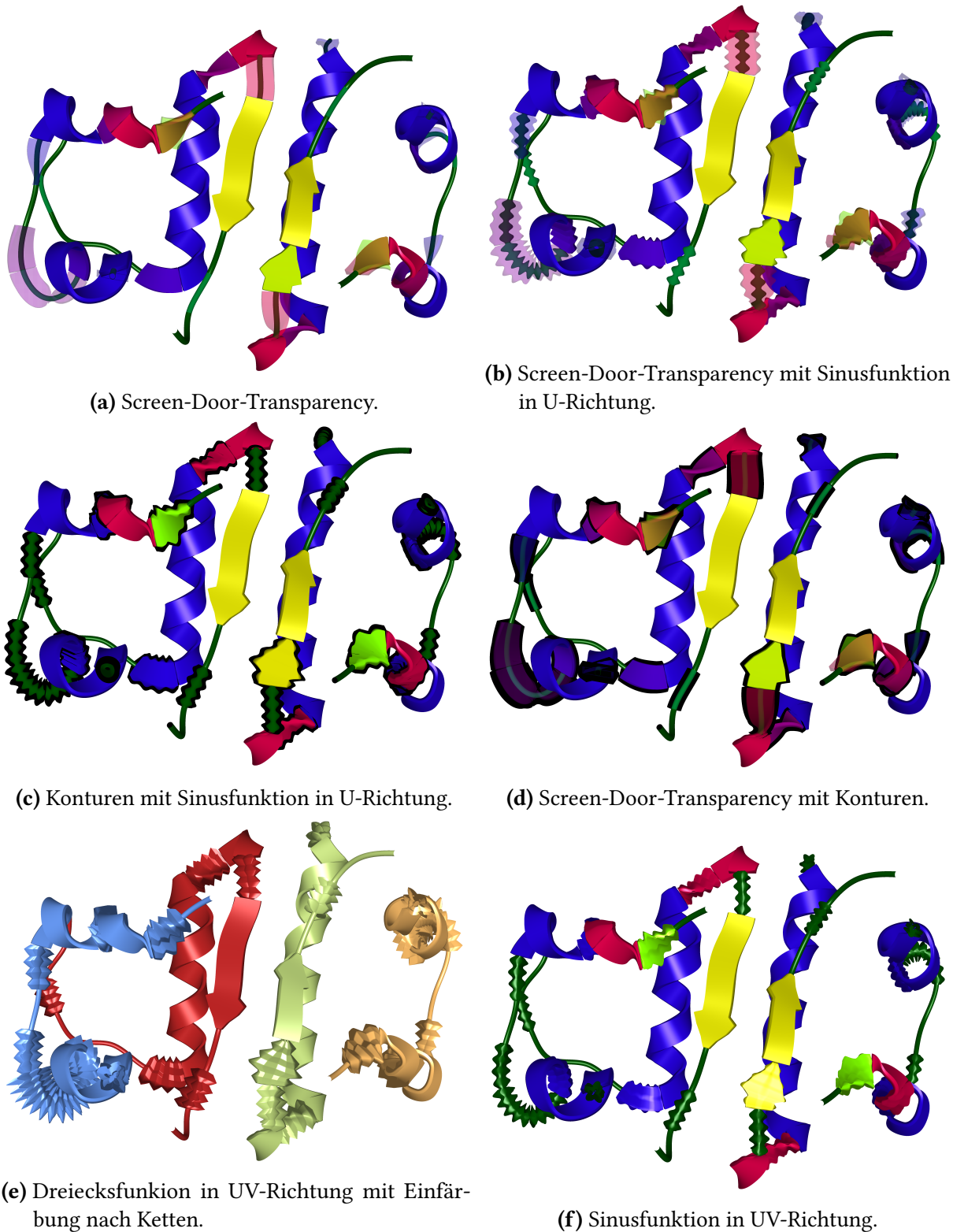
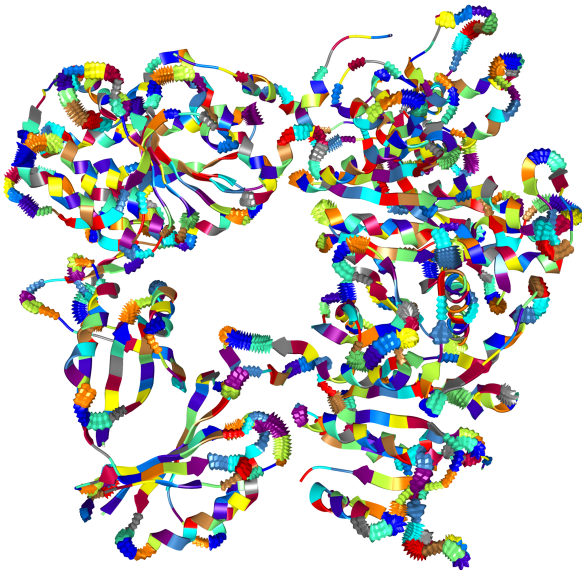
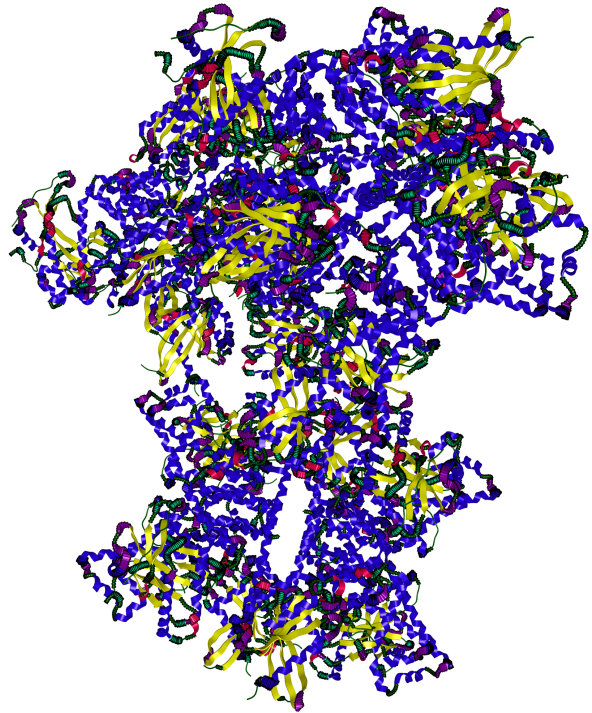


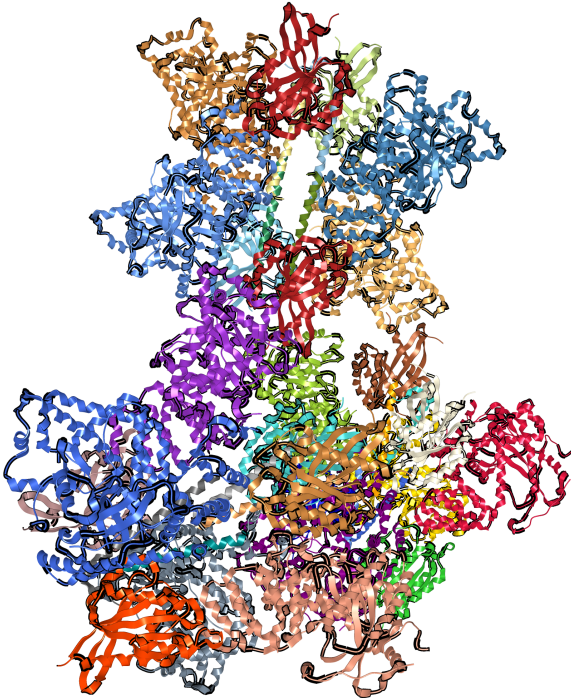
Abbildung 6.5: 1RWE: Cartoon-Darstellungen. (2)



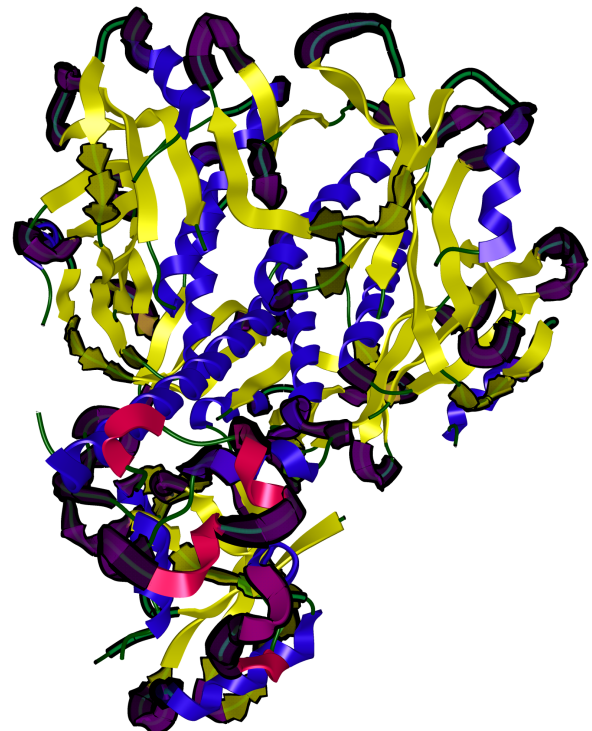
(a) 2JJ4: Dreiecksfunktion in UV-Richtung mit unterschiedlicher Einfärbung der Aminosäuren.



(b) 4D0M: Konturen mit Dreiecksfunktion in U-Richtung.



(c) 4D0M: Konturen mit Einfärbung nach Ketten.



(d) 1TII: Screen-Door-Transparency mit Konturen.

Abbildung 6.6: Cartoon-Darstellungen.

6.1.3 Aufgeteilte Ansicht

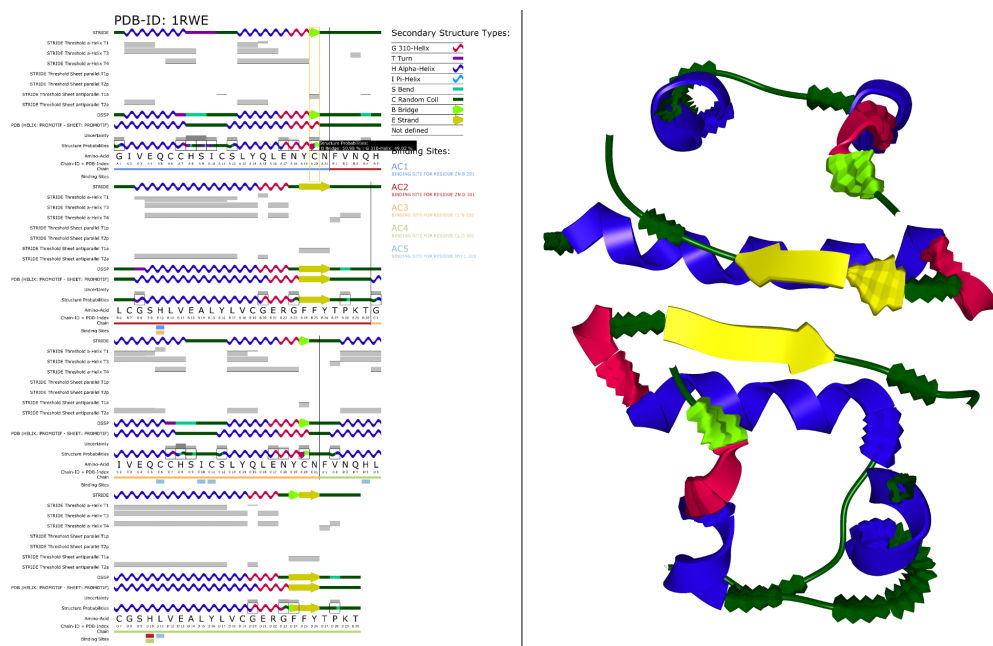


Abbildung 6.7: 1RWE: Aufgeteilte Ansicht.

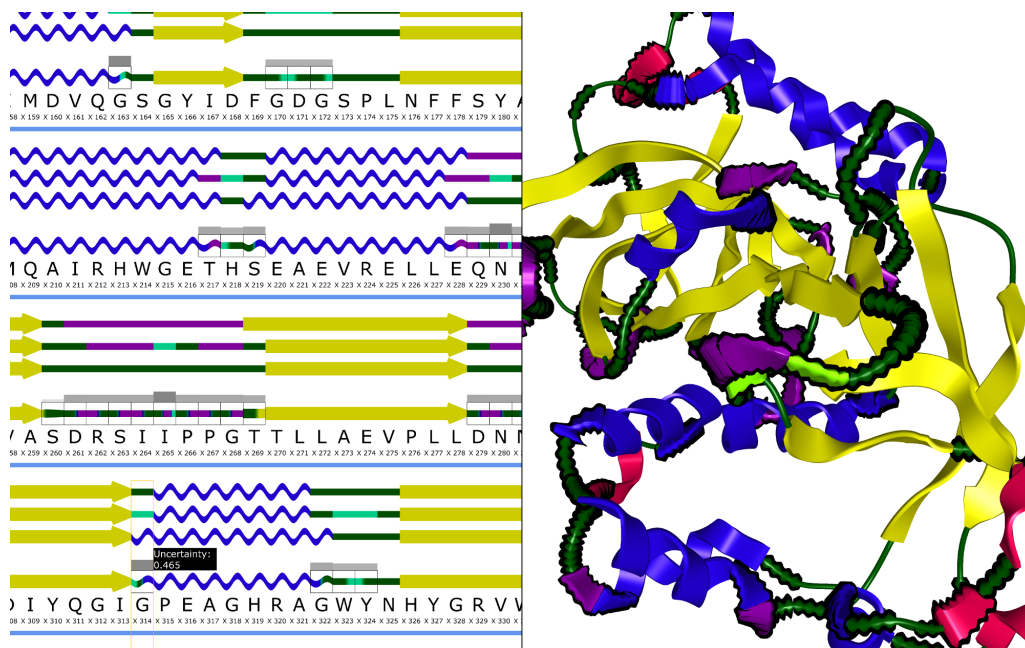


Abbildung 6.8: 2AE0: Aufgeteilte Ansicht.

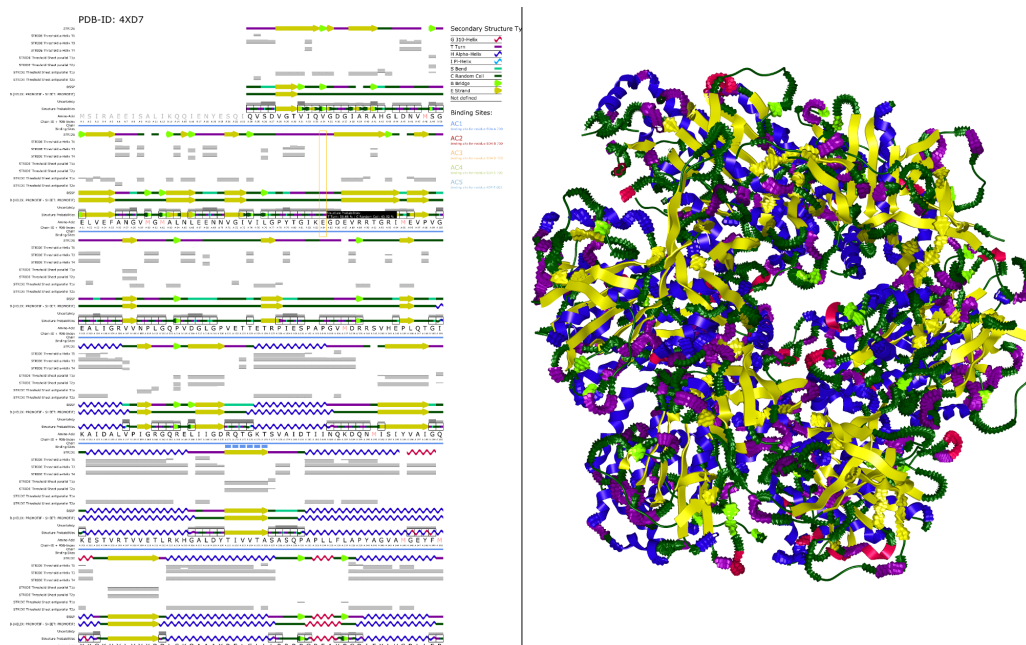


Abbildung 6.9: 4XD7: Aufgeteilte Ansicht.

6.2 Bewertung und Ausblick

In diesem Abschnitt werden die Ergebnisse der hier vorliegenden Arbeit diskutiert. An den entsprechenden Stellen wird außerdem auf mögliche Verbesserungen und Erweiterungen eingegangen. Es wird weder auf die Erweiterung um ein zusätzliches Verfahren zu dessen Qualitätsanalyse noch auf die Möglichkeit, die Stabilität eines Sekundärstrukturelements in dynamischen Daten zu untersuchen, eingegangen, da eine Umsetzung den zeitlichen Rahmen dieser Arbeit gesprengt hätte.

6.2.1 Performanz

Auch wenn der Fokus bei der Implementierung in erster Linie nicht auf der Optimierung der Performanz lag, sollte die Interaktivität der Darstellung jedoch erhalten bleiben. Als Maß der Performanz dient die Anzahl der pro Sekunde gezeichneten Bilder, welche in *fps* (frames per second) angegeben werden. Die den Tests zugrunde liegende Hardware besitzt folgende Eigenschaften:

Computer:	Intel Quad-Core i7-2600 - 3,4 GHz - 16 GB RAM
Betriebssystem:	Windows 10 Enterprise 64bit
Grafikkarte:	NVIDIA GeForce GTX 660 Ti (Mittelklasse)
Display-Auflösung:	1920 × 1200 Pixel

Die Untersuchung erfolgte für die vier Proteine 1RWE, 1TII, 4XD7 und 4D0M in Tabelle 6.1. Die verfügbaren Parameter mit Einfluss auf die Performanz sind die Anzahl der Dreieck-Primitiven für jede der beiden Ansichten. Im Sequenz-Diagramm kann die Anzahl der Dreiecke der einzelnen, pro Kachel gezeichneten, Strukturformen angegeben werden. In der Cartoon-Darstellung verhält sich die Anzahl der Dreiecke quadratisch zu der pro Aminosäure definierbaren Unterteilungsstufe der Strukturformen. Die hier verwendeten Unterteilungsstufen 16, 32 und 64 gelten für beide Ansichten. Zusätzlich hängt die Performanz von der Anzahl der Geometriepässe ab, welche sich anhand der gewählten Darstellung bzw. aus der Kombination mehrerer Unsicherheitsdarstellungen in der Cartoon-Darstellung ergeben.

PDB-ID	#AMS	Unterteilungsstufen der Geometrie								
		16	32	64	16	32	64	16	32	64
		GVZ (1)			GVZ+KNT (2)			GVZ+KNT+SDT (5)		
1RWE	102	263	232	99	210	145	52	153	87	27
1TII	738	48	34	15	39	23	8	28	13	4
4XD7	3373	10	7	3	8	5	2	5	3	1
4D0M	9996	4	2	1	< 1	< 1	< 1	< 1	< 1	* < 1

Tabelle 6.1: Es sind die *fps* für drei unterschiedliche Kombinationen von Darstellungen und für je drei Unterteilungsstufen der Geometrie (16, 32 und 64) angegeben. #AMS = Anzahl der Aminosäuren, GVZ = Geometrieverzerrung, KNT = Konturen, SDT = Screen-Door-Transparency; die Anzahl der Geometriepässe ist in Klammern hinter den Verfahren angegeben. * Bei Darstellung in der Cartoon-Darstellung ohne Sequenz-Diagramm sind es noch ≈ 5 fps.

Wie anhand der in Tabelle 6.1 angegebenen *fps* zu sehen ist, können die Darstellungen für Proteine mit einer Größe bis zu 3000 Aminosäuren über eine entsprechende Verringerung der Geometriestufe interaktiv dargestellt werden. Für durchschnittlich große Aminosäuren ist die Interaktivität auf jeden Fall gegeben. Bei größeren Proteinen bietet die Anpassung der Unterteilungsstufen eine effektive Möglichkeit, sich die Unsicherheit noch interaktiv anzeigen zu lassen. Zumal auch bei einer geringeren Unterteilungsstufe der Detailgrad der Geometrie für eine gute Interpretation der Unsicherheit noch ausreichend hoch ist. Eine weitere Absenkung der Unterteilung hätte jedoch eine ungenauere geometrische Repräsentation der Unsicherheitsdarstellung zur Folge. Dies würde eine zusätzliche, von der Darstellung verursachte, Visualisierungsunsicherheit erzeugen und somit die Aussagekraft der dargestellten Unsicherheit beeinträchtigen. Das jeweils auffällige Abfallen der *fps* bei der Erhöhung der Unterteilung von 32 auf 64 lässt sich damit erklären, dass sich bei einer Verdoppelung der Unterteilungsstufe sich die Anzahl der Dreiecke vervierfacht. Das Problem der Performanzreduktion ließe sich zum einen durch eine Optimierung der Darstellungsverfahren abmildern. Auch die Wahl anderer entsprechend performanter implementierbare Methoden wäre eine Möglichkeit.

6.2.2 Unsicherheitsmodell

Die Berücksichtigung der Unsicherheit, welche bei STRIDE aus den empirisch festgelegten Schwellenwerten entsteht, erfolgt anhand der Definition von Vertrauensintervallen jeweils um die Schwellenwerte. Die Strukturtyp-Wahrscheinlichkeiten werden dann durch die Überprüfung der Zuordnungskriterien für Werte innerhalb der Vertrauensintervalle entsprechend angepasst. Diese hier entwickelte Vorgehensweise lässt sich auch auf beliebige andere Verfahren anwenden, welche ihre Zuordnungskriterien über Schwellenwerte definieren (z.B. DSSP oder PROSIGN).

Die Darstellung der Schwellenwerte für die berechneten Werte erfolgt über ein um den Schwellenwert angeordnetes Säulendiagramm. Für jeden der sieben Schwellenwerte von STRIDE wird zusätzlich eine Zeile im Sequenz-Diagramm eingeblendet. Die Einblendung der Schwellenwerte mindert die Vergleichbarkeit der unterschiedlichen Zuordnungen. Sie eignet sich nur dann, wenn tatsächlich anhand der Schwellenwerte deren Einfluss auf die resultierende Zuordnung von STRIDE untersucht werden soll. Dafür ist allerdings eine detaillierte Kenntnis der Zuordnungskriterien von STRIDE notwendig. Die Werte der veränderten Strukturtyp-Wahrscheinlichkeiten sind nur über den Tooltip der Sekundärstruktur-Zeile von STRIDE im Sequenz-Diagramm ersichtlich. Für eine unmittelbare Interpretation wäre eine analoge Darstellung zu den Strukturtyp-Wahrscheinlichkeiten auf der Ebene der Aminosäuren hilfreich. Anhand der Häufigkeit, mit welcher die Zuordnungsunsicherheit von STRIDE die resultierende Unsicherheit an einer Aminosäure verändert, wurde der Einfluss des Schwellenwert-Kriteriums abgeschätzt. Dafür wurden die Proteine 1RWE, 1TII, 4XD7, 2AE0, 2jj4, 1IGT, 4D0M und 4UER mit insgesamt 26799 Aminosäuren untersucht. Die Strukturtyp-Wahrscheinlichkeiten von α -Helices werden demnach in 0,92% der Fälle, gemessen an der Gesamtzahl der hier berücksichtigten Aminosäuren, verändert. Bei β -Brücken und β -Strängen sind es zusammen genommen 0,54%. Daraus ergibt sich, dass die Strukturtyp-Wahrscheinlichkeiten bei 1,46% der Aminosäuren durch die Zuordnungsunsicherheit von STRIDE beeinflusst werden. Da Veränderungen fast ausschließlich an den Enden von Sekundärstrukturen erfolgen, erhöht sich dieser Einfluss noch, wenn zusätzlich Strukturängen berücksichtigt werden. Es lässt sich somit zum einen ableiten, dass das gewählte Vertrauensniveau für die Schwellenwerte von 95% nicht zu gering gewählt wurde und der Einfluss nicht unverhältnismäßig hoch ist. Zum anderen ist der Einfluss auch nicht zu gering, in welchem Fall die Zuordnungsunsicherheit von STRIDE einen zu vernachlässigbaren Einfluss hätte.

Die pro Aminosäure berechneten Unsicherheitswerte werden anhand von Säulen-Diagrammen im Sequenz-Diagramm dargestellt. Um die konkreten Werte zu ergänzen, steht ein Tooltip zur Verfügung. Da diese Werte allerdings immer nur für eine Aminosäure angezeigt werden, erfolgt der visuelle Vergleich der Unsicherheitswerte hauptsächlich über die dafür verwendeten visuellen Variablen *Helligkeit* und *Breite* (bzw. *Höhe*). Sowohl das relative als auch das absolute Verhältnis der Unsicherheiten zwischen den Aminosäuren sind damit leicht abzulesen. Dies gilt bei der Betrachtung kleiner Sequenzausschnitte gleichermaßen wie für die Analyse einer größerer Sequenz eines Proteins.

Der Einfluss der Diskrepanzwerte auf die Strukturtyp-Wahrscheinlichkeiten ist in Bezug auf die Größenordnung angemessen und plausibel. Insgesamt wird das hier entworfene Modell der Unsicherheit als umfassend und schlüssig beurteilt. Es kann als Grundlage für weitere wissenschaftliche Arbeiten dienen. Es ist analog zur hier beschriebenen Vorgehensweise einfach um weitere Verfahren erweiterbar. Die Anzahl der Diskrepanz-Matrizen steigt dabei annähernd quadratisch zur Anzahl der berücksichtigten Verfahren.

Eine denkbare Erweiterung wäre z.B., die Werte der Diskrepanz-Matrizen anhand der verfügbaren manuellen Zuordnungen zu berechnen statt sie manuell zu bestimmen. Dies könnte durch die Filterung der manuellen Zuordnung anhand eines probabilistischen Entscheidungsbaums erfolgen. Mittels eines Algorithmus zur statistischen Klassifikation könnten die Werte der Diskrepanz-Matrizen dann automatisiert berechnet werden.

6.2.3 Evaluierung der Visualisierungen

Die folgenden Bewertungen der hier verwendeten Darstellungen der Unsicherheit beruhen ausschließlich auf der eigenen Beurteilung. Dadurch, dass mir die genaue Implementierung bekannt ist und ich weiß, was die einzelnen Darstellungen aussagen sollten, kann nur eine voreingenommene Evaluation gegeben werden. Bei der Analyse geeigneter Darstellungen in Abschnitt 5.1 wurde die Kritik angeführt, dass es für Unsicherheitsvisualisierungen zu wenig ausreichend objektive Evaluationen gibt. Auch im Rahmen dieser Arbeit konnte eine solche Evaluation anhand einer Benutzerstudie oder anhand einer umfassenden Analyse mit den theoretischen Grundsätzen zur Wahrnehmung (z.B. nach Tufte, Bertin oder Ware, siehe Abschnitt 5.1) nicht erfolgen. Deshalb ist es für eine objektive Beurteilung notwendig, diesen Schritt auch für diese Arbeit nachzuholen. Es wird sich in Zukunft zeigen, wie andere Benutzer die Darstellungen interpretieren und ob die hier gegebene Einschätzung bestätigt wird. Unsicherheitsdarstellungen können Entscheidungen auf der Grundlage der dargestellten Daten sinnvoll ergänzen und erleichtern. Voraussetzung dafür ist sowohl ein ausreichendes Wissen über die Art und das Zustandekommen der Daten als auch eine gewisse Kenntnis der Methode, welche zur Darstellung der Unsicherheit verwendet wurde. Während bei den Darstellungen der Unsicherheit im Sequenz-Diagramm der quantitative Aspekt im Vordergrund steht, ist es bei der Cartoon-Darstellung der qualitative. Um die Ergänzung der beiden Ansichten zu vervollständigen, müssten sich jeweils die gleichen Aminosäuren über eine Selektion und Hervorhebung verknüpfen lassen.

Sequenz-Diagramm

Während im Abschnitt 6.2.2 die Darstellung der Unsicherheits- und Schwellenwerte bewertet wurde, liegt in diesem der Fokus auf der Darstellung der Strukturtyp-Wahrscheinlichkeiten. Da sich die Strukturformen geometrisch deutlich voneinander unterscheiden, sind über das

Struktur-Morphing ausgedrückte unterschiedliche Strukturtyp-Wahrscheinlichkeiten qualitativ gut zu differenzieren. Allerdings mit der Einschränkung, dass auf diese Weise nur die in den Strukturformen zusammengefassten Gruppen von Strukturtypen unterschieden werden können. Durch zusätzliche Strukturgeometrien, beispielsweise für Umdrehungen, könnte die Aussagekraft des Struktur-Morphings zusätzlich erhöht werden.

Die Interpretation der interpolierten Farben ist nicht intuitiv. Es sollte deshalb bekannt sein, dass die Anzahl der unterschiedlichen Farben innerhalb des Intervall, in welchem die Strukturfarben interpoliert werden, proportional zur Diskrepanz zwischen farblich benachbart dargestellten Strukturtypen ist. Je mehr unterschiedliche Strukturfarben innerhalb einer Kachel zu sehen sind, desto größer ist die Diskrepanz zwischen den dort wahrscheinlichen Strukturtypen. So wird innerhalb der einzelnen Kacheln für die Strukturtyp-Wahrscheinlichkeiten zusätzlich die Information über die Diskrepanz der Strukturtypen dargestellt. Der ursprüngliche quantitative Zusammenhang zwischen den Strukturfarben außerhalb der Interpolationsintervalle und den entsprechenden Strukturtyp-Wahrscheinlichkeiten bleibt erhalten. Unter der Voraussetzung, dass die Interpretationsweise bekannt ist, wird diese Darstellung als geeignet und hilfreich eingeschätzt. An dieser Darstellung wird ersichtlich, dass für eine intuitive Wahrnehmung einer Unsicherheitsdarstellung die Einfachheit im Vordergrund stehen muss. Umgekehrt lässt sich daraus schließen, dass für kompliziertere Darstellungen das Angebot einer ausreichenden Erklärung mitberücksichtigt werden muss.

Die Möglichkeit, die sicheren Zuordnungen in der Eigenschaftszeile der Strukturtyp-Wahrscheinlichkeiten auszublenden, eignet sich dafür, den Fokus auf die unsicheren Zuordnungen zu lenken. Die zusätzlichen Umrandungen solcher Kacheln unterstützt die Differenzierung zusätzlich. Die textuelle Erweiterung um die konkreten Werte der Strukturtyp-Wahrscheinlichkeiten in Form des Tooltips ergänzt den quantitativen Aspekt des Sequenz-Diagramms sinnvoll. Die zusätzlich einblendbaren Eigenschaftszeilen für die Darstellung der Unsicherheit überschneiden sich nicht mit der ursprünglichen Möglichkeit, die Daten zu analysieren, sondern können ergänzend parallel dazu betrachtet werden.

Als eine weitere potentielle Darstellung im Sequenz-Diagramm wurden ursprünglich drei Animationsarten der Geometrie und Farbe, ausgehend von den Strukturtyp-Wahrscheinlichkeiten, untersucht. Aufgrund einer visuellen Überbeanspruchung der Benutzer schied dieser Ansatz in den weiteren Betrachtungen jedoch aus. Auch eine Repräsentation der Strukturtyp-Wahrscheinlichkeiten mittels Glyphen wurde als mögliche Darstellung geprüft. Es konnte jedoch keine zufriedenstellende Repräsentation gefunden werden. Allerdings wäre eine Darstellung der Unsicherheit über geeignete Glyphen durchaus vorstellbar.

Cartoon-Darstellung

Eine der drei in der Cartoon-Darstellung zur Verfügung stehenden Darstellungen für die Unsicherheit ist die Geometrieverzerrung der Strukturformen von unsicheren Sekundärstruktur-Zuordnungen. Die Verschiebung der Oberfläche ist dabei sowohl entlang einer Sinus-Funktion

als auch entsprechend einer Dreieck-Funktion möglich. Die Funktionen können entweder in Richtung der Hauptkette (U-Richtung) angewendet werden oder zusätzlich parallel dazu (UV-Richtung). Der Einfluss der Geometrieverzerrung ist stark von den gewählten Skalierungsfaktoren für die Amplitude und die Frequenz anhängig. Es wird empfohlen, die Skalierung der Frequenz auf 1 und die der Amplitude im Bereich 1-3 festzulegen. Die zugrundeliegenden Strukturformen sind damit noch gut sichtbar und die Aminosäuren mit unsicheren Zuordnungen heben sich auch bei größeren Proteinen gut erkennbar von den sicheren Zuordnungen ab. Generell ist eine quantitative Aussage anhand der Geometrieverzerrung schwer. Bei der Anwendung in U-Richtung können quantitative Unterschiede jedoch differenziert werden.

Bei der alternativen Darstellung der Unsicherheit über Konturen spielt auch der Skalierungsfaktor wieder eine entscheidende Rolle. Für eine angemessene Darstellung der Konturen muss ein guter Mittelwert zwischen den Extremen der visuellen Überbeanspruchung und einer nicht mehr gegebenen Wahrnehmbarkeit der Konturen gewählt werden. Ein solcher Skalierungsfaktor liegt im Bereich 6-10. Da die Konturen an der Außenseite der Strukturgeometrie liegen, wird die Wahrnehmung der Sekundärstruktur und die Analyse anderer Eigenschaften nicht gestört. Die Konturen erhöhen zudem den Kontrast zwischen der im Vordergrund stehenden und der im Hintergrund liegenden Struktur eines Proteins. Dies erleichtert vor allem bei größeren Proteinen die Wahrnehmung der unsicheren Bereiche. Auch bei den Konturen ist die Aussagekraft hauptsächlich qualitativ. Die Veränderung der Konturenfarbe wurde getestet. Die Überschneidung mit der für andere Eigenschaften verwendeten visuellen Variablen *Farbe* ist dabei groß. Deshalb eignet sich diese Option wie vermutet nicht.

Die Visualisierung der Strukturtyp-Wahrscheinlichkeiten erfolgt mittels Screen-Door-Transparency über die Methode des Ditherings. Bei maximal vier Strukturtypen pro Aminosäure ist die Wahrscheinlichkeit größer als Null. Dies entspricht somit auch der maximalen Anzahl an gleichzeitig je Aminosäure prozentual übereinander gezeichneten Strukturformen in den jeweiligen Strukturfarben. Die Differenzierung von mehr als zwei Strukturfarben ist dabei nur in der Betrachtung kleiner Sequenzabschnitte gut möglich. Die quantitative Abschätzung der Strukturtyp-Wahrscheinlichkeit ist auch anhand der Körnigkeit nur grob möglich. Dies könnte durch die Verwendung anderer Dither-Matrizen verbessert werden. Bei einer kleineren 4×4 Bayer-Matrix ist beispielsweise die Musterbildung durch deren häufigere Wiederholung im Bildraum ausgeprägter. Außerdem würden nur 16 Werte unterschieden, entgegen der 64 bei der hier verwendeten 8×8 Bayer-Matrix. Dies könnte auf der einen Seite die Wahrnehmung von sich unterscheidenden Strukturtyp-Wahrscheinlichkeiten erhöhen. Auf der anderen Seite würde durch diese drastische Reduktion der unterschiedenen Strukturtyp-Wahrscheinlichkeiten eine zusätzliche Unsicherheit, hervorgerufen durch die Visualisierung, entstehen. Die Unterscheidbarkeit der über das Dithering kombinierten Strukturfarben wird durch das Weglassen des spekularen Anteil des Beleuchtungsmodells signifikant erhöht. Ohne die Einfärbung der Cartoon-Darstellung des Proteins in den Strukturfarben können die Strukturformen ausschließlich anhand der Körnigkeit quantitativ abgeschätzt werden.

Die Kombination von Geometrieverzerrung und Konturen führt schnell zu einer visuellen Überbeanspruchung, was durch eine Reduktion der Skalierungsfaktoren nur bedingt beeinflusst

werden kann. Wird die Geometrieverzerrung nur in U-Richtung gewählt, ist eine gegenseitige Ergänzung durch die jeweils verwendete visuelle Variable *Breite* allerdings durchaus gegeben. Wird die Darstellung der Screen-Door-Transparency entweder mit der Geometrieverzerrung oder den Konturen kombiniert, profitiert die Darstellung der Strukturtyp-Wahrscheinlichkeiten durch eine bessere Unterscheidbarkeit zu sicheren Zuordnungen. Die Kombination aller drei Darstellungen führt zu einer visuellen Überbeanspruchung und eignet sich daher nicht.

Bei der hier berechneten Unsicherheit und den Strukturtyp-Wahrscheinlichkeiten handelt es sich um relative Aussagen. Es besteht kein absoluter Zusammenhang zur dargestellten Sekundärstruktur. Dieser relative Bezug ermöglicht es, über Skalierungsfaktoren das Verhältnis der Unsicherheitsdarstellung zur gesamten Visualisierung des Proteins nach eigenem Ermessen festzulegen.

Eine weitere Möglichkeit zur Visualisierung von Unsicherheit in der Cartoon-Darstellung wäre das sogenannte *Stippling* [Sec02], ein weiterer auf Punkten basierender Ansatz. Interessant wäre zudem eine Untersuchung, inwiefern sich unterschiedliche Beleuchtungsmodelle zur Darstellung von Unsicherheit eignen. Auch eine andere Vorgehensweise bei der Veränderung der Geometrie oder dreidimensionales Struktur-Morphing könnten sich als Darstellungen von Unsicherheit in der dreidimensionalen Cartoon-Darstellung eignen.

Kapitel 7

Zusammenfassung

Im Rahmen dieser Diplomarbeit wurde ein Modell der Unsicherheit für abweichende Sekundärstruktur-Zuordnungen entwickelt. Ausgangspunkt dafür waren die Verfahren STRIDE, DSSP und die in den PDB-Dateien enthaltenen Angaben zur Sekundärstruktur von PROMOTIF oder der manuellen Zuordnung. Auf der Betrachtungsebene der einzelnen Verfahren wurde für die Zuordnungsunsicherheit von STRIDE ein Schwellenwert-Kriterium definiert. Die Zuordnungskriterien für α -Helices und β -Stränge werden bei STRIDE anhand empirisch festgelegter Schwellenwerte festgelegt. Um diese Schwellenwerte herum wurden Vertrauensintervalle definiert, innerhalb welcher eine Modifikation der Zuordnungswahrscheinlichkeiten erfolgt. Auf der Ebene der einzelnen Aminosäuren erfolgte der Vergleich zwischen den Zuordnungen der Verfahren. Mittels dafür definierter Diskrepanz-Matrizen wurde für jeden Strukturtyp eine Zuordnungswahrscheinlichkeit abgeleitet. Dadurch ergibt sich an jeder Aminosäure eine Wahrscheinlichkeitsverteilung über die möglichen Strukturtypen. Ausgehend davon wurde die Unsicherheit über die Bestimmung der Standard-Abweichung auf einen Wert je Aminosäure reduziert. Basierend auf einer Analyse bestehender Visualisierungen für Unsicherheit wurden für das zweidimensionale Sequenz-Diagramm und die dreidimensionale Cartoon-Darstellung jeweils mehrere unterschiedliche Darstellungen ausgewählt. Die Strukturtyp-Wahrscheinlichkeiten wurden im Sequenz-Diagramm mittels Struktur-Morphing und Farbinterpolation im HSL-Farbraum dargestellt, die Unsicherheits- und Schwellenwerte über Säulen-Diagramme. In der Cartoon-Darstellung wurde der Unsicherheitswert in Form von Geometreiverzerrung oder Konturen dargestellt. Die Repräsentation der Strukturtyp-Wahrscheinlichkeiten erfolgte über Screen-Door-Transparency. Die Unsicherheit der Sekundärstruktur-Zuordnung steht den Benutzern somit in beiden Ansichten eines Proteins als zusätzliche Information zur Verfügung. Die beiden Betrachtungsweisen eines Proteins können gleichzeitig nebeneinander angezeigt werden. Dies ermöglicht die jeweilige Korrelation und Ergänzung der unterschiedlichen Unsicherheitsdarstellungen. Es wird eine vollständigere und genauere Wiedergabe der vorhandenen Daten bezüglich der Sekundärstruktur-Zuordnung erreicht.

Abbildungsverzeichnis

2.1	Visualisierungspipeline der Unsicherheit. Quelle: Abbildung basiert auf [PWL97]	9
3.1	Unterschiedliche Struktur- und Sequenz-Visualisierungen.	13
3.2	Verschiedene Visualisierungen von Unsicherheit.	15
3.3	Visualisierungen von Unsicherheit bei Molekülen.	16
4.1	Normalverteilungen für Werte, welche ein Kriterium erfüllen bzw. nicht erfüllen, und das resultierende Intervall um den Schwellenwert T_i	31
4.2	Beispiel, wie sich die Zuordnungsunsicherheit bei STRIDE auswirkt.	35
4.3	Veranschaulichung der Diskrepanzen zwischen den Strukturtypen.	40
5.1	Vollständiges Sequenz-Diagramm für das Protein mit der PDB-ID 1RWE.	51
5.2	Unterteilung der unterschiedenen Strukturformen in Dreiecksprimitiven. Von links oben nach rechts unten: Sinus-Kurve für Helices; schmaler Balken für Biegung, Umdrehung und zufällige Windung; breiter Balken für β -Strang; Pfeilspitze für β -Brücken und letzte Aminosäure eines β -Strangs.	52
5.3	Interpolation zweier Strukturfarben C1 und C2, je für die Strukturtyp-Wahrscheinlichkeiten $p_V(s_{[v][a]})$ und $p_V(t_{[v][a]})$ innerhalb des Intervalls b	56
5.4	Links: Sinusfunktion. Rechts: Dreiecksfunktion.	59
5.5	Berechnung der Normalenabweichung Δt in die Richtung von \vec{t} (Erklärung siehe Text).	60
6.1	1RWE: Vollständiges Sequenz-Diagramm.	66
6.2	1RWE: Vergrößerter Ausschnitt aus dem Sequenz-Diagramm.	66
6.3	1IGT: Ausschnitt aus dem Sequenz-Diagramm, die Schwellenwerte von STRIDE sind ausgeblendet.	67
6.4	1RWE: Cartoon-Darstellungen. (1)	67
6.5	1RWE: Cartoon-Darstellungen. (2)	68
6.6	Cartoon-Darstellungen.	69
6.7	1RWE: Aufgeteilte Ansicht.	70
6.8	2AE0: Aufgeteilte Ansicht.	70
6.9	4XD7: Aufgeteilte Ansicht.	71

7.1	Zusammenfassung der entwickelten Darstellungen: Aufgeteilte Ansicht für das Protein 1TII.	80
-----	---	----

Tabellenverzeichnis

4.1	Die jeweils von den Verfahren unterschiedenen Sekundärstrukturtypen. . . .	29
4.2	Schwellenwert-Intervalle für ein einseitiges Vertrauensniveau von $p_T = 95\%$.	32
5.1	Die Strukturfarben in HSL- und RGB-Werten und in Kombination mit den Strukturformen.	55
6.1	Es sind die f_{ps} für drei unterschiedliche Kombinationen von Darstellungen und für je drei Unterteilungsstufen der Geometrie (16, 32 und 64) angegeben. . . .	72

- [CET+93] N. Colloc'h, C. Etchebest, E. Thoreau, B. Henrissat, J. P. Mornon. „Comparison of three algorithms for the assignment of secondary structure in proteins: The advantages of a consensus assignment“. In: *Protein Eng. Des. Sel.* 6.4 (1993), S. 377–382. DOI: [10.1093/protein/6.4.377](https://doi.org/10.1093/protein/6.4.377) (zitiert auf S. 12).
- [CG14] M. Correll, M. Gleicher. „Error bars considered harmful: Exploring alternate encodings for mean and error“. In: *IEEE Trans. Vis. Comput. Graph.* 20.12 (2014), S. 2142–2151. DOI: [10.1109/TVCG.2014.2346298](https://doi.org/10.1109/TVCG.2014.2346298) (zitiert auf S. 14).
- [DRSR15] S. M. Dabdoub, W. Rumpf, A. D. Shindhelm, W. C. Ray. „MoFlow: visualizing conformational changes in molecules as molecular flow improves understanding“. In: *BMC Proc.* 9.6 (2015), S. 5. DOI: [10.1186/1753-6561-9-S6-S5](https://doi.org/10.1186/1753-6561-9-S6-S5) (zitiert auf S. 16, 17).
- [FA95] D. Frishman, P. Argos. „Knowledge-based protein secondary structure assignment“. In: *Proteins Struct. Funct. Genet.* 23.4 (1995), S. 566–579. DOI: [10.1002/prot.340230412](https://doi.org/10.1002/prot.340230412) (zitiert auf S. 2, 7, 11, 25, 30, 33, 38, 40).
- [FBW16] F. Ferstl, K. Bürger, R. Westermann. „Streamline Variability Plots for Characterizing the Uncertainty in Vector Field Ensembles“. In: *IEEE Trans. Vis. Comput. Graph.* 22.1 (2016), S. 767–776. DOI: [10.1109/TVCG.2015.2467204](https://doi.org/10.1109/TVCG.2015.2467204) (zitiert auf S. 14, 15).
- [FDFH95] J. D. Foley, A. van Dam, S. K. Feiner, J. F. Hughes. *Computer Graphics: Principles and Practice in C*. 2. Aufl. Addison-Wesley Professional, 1995, S. 1200. ISBN: 0201848406 (zitiert auf S. 62).
- [Fie76] R. Fieldman. „Atlas of Macromolecular structure on Microfiche (AMSOM)“. In: *Document* 13.1.1 (1976), S. 1 (zitiert auf S. 43).
- [FRZ99] K. Fidelis, B. Rost, A. Zemla. „A Modified Definition of Sov, a Segment-Based Measure for Protein Secondary Structure Prediction Assessment 1“. In: 223.8 (1999), S. 220–223 (zitiert auf S. 11).
- [Geo09] H.-O. Georgii. *Stochastik. Einführung in die Wahrscheinlichkeitstheorie und Statistik*. 4. Aufl. Berlin: Walter de Gruyter, 2009, S. 416–421. ISBN: 978-3-11-021526-7. DOI: [10.1515/9783110215274](https://doi.org/10.1515/9783110215274) (zitiert auf S. 32).
- [Ger98] N. Gershon. „Visualization of an imperfect world“. In: *IEEE Comput. Graph. Appl.* 18.4 (1998), S. 43–45. DOI: [10.1109/38.689662](https://doi.org/10.1109/38.689662) (zitiert auf S. 14).
- [GHL15] H. Guo, J. Huang, D. H. Laidlaw. „Representing Uncertainty in Graph Edges: An Evaluation of Paired Visual Variables“. In: *IEEE Trans. Vis. Comput. Graph.* 21.10 (2015), S. 1173–1186. DOI: [10.1109/TVCG.2015.2424872](https://doi.org/10.1109/TVCG.2015.2424872) (zitiert auf S. 14, 48).
- [GKM+15] S. Grottel, M. Krone, C. Müller, G. Reina, T. Ertl. „MegaMol - A prototyping framework for particle-based visualization“. In: *IEEE Trans. Vis. Comput. Graph.* 21.2 (2015), S. 201–214. DOI: [10.1109/TVCG.2014.2350479](https://doi.org/10.1109/TVCG.2014.2350479) (zitiert auf S. 2).

- [GR02] G. Grigoryan, P. Rheingans. „Probabilistic surfaces: point based primitives to show surface uncertainty“. In: *IEEE Vis. 2002. VIS 2002.* (2002), S. 147–153. DOI: [10.1109/VISUAL.2002.1183769](https://doi.org/10.1109/VISUAL.2002.1183769) (zitiert auf S. 15).
- [GR04] G. Grigoryan, P. Rheingans. „Point-based probabilistic surfaces to show surface uncertainty“. In: *IEEE Trans. Vis. Comput. Graph.* 10.5 (2004), S. 564–573. DOI: [10.1109/TVCG.2004.30](https://doi.org/10.1109/TVCG.2004.30) (zitiert auf S. 15, 16, 49).
- [GS06] H. Griethe, H. Schumann. „The Visualization of Uncertain Data: Methods and Problems“. In: *Proc. Simul. und Vis. SIMVIS '06 vi.JANUARY 2006* (2006), S. 143–156 (zitiert auf S. 48).
- [HF04] M. Heinig, D. Frishman. „STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins“. In: *Nucleic Acids Res.* 32.WEB SERVER ISS. (2004), S. 500–502. DOI: [10.1093/nar/gkh429](https://doi.org/10.1093/nar/gkh429). URL: <http://webclu.bio.wzw.tum.de/stride/> (zitiert auf S. 26).
- [HG93] G. J. Hunter, M. F. Goodchild. *Managing uncertainty in spatial databases: Putting theory into practice.* 1993 (zitiert auf S. 8).
- [HGVV15] P. Hermosilla, V. Guallar, A. Vinacua, P. P. Vázquez. „Instant Visualization of Secondary Structures of Molecular Models“. In: *Proc. Eurographics Work. Vis. Comput. Biol. Med.* (2015), S. 51–60. DOI: [10.2312/vcbm.20151208](https://doi.org/10.2312/vcbm.20151208) (zitiert auf S. 58).
- [HSP+08] S. R. Hosseini, M. Sadeghi, H. Pezeshk, C. Eslahchi, M. Habibi. „PROSIGN: A method for protein secondary structure assignment based on three-dimensional coordinates of consecutive C alpha atoms“. In: *Comput. Biol. Chem.* 32.6 (2008), S. 406–411. DOI: [10.1016/j.compbiolchem.2008.07.027](https://doi.org/10.1016/j.compbiolchem.2008.07.027) (zitiert auf S. 2, 11).
- [HT96] E. G. Hutchinson, J. M. Thornton. „PROMOTIF—a program to identify and analyze structural motifs in proteins.“ In: *Protein Sci.* 5.2 (1996), S. 212–220. DOI: [10.1002/pro.5560050204](https://doi.org/10.1002/pro.5560050204) (zitiert auf S. 7, 38, 42).
- [IUP74] IUPAC-IUB Commission on Biochemical Nomenclature (CBN). „Abbreviations and symbols for description of conformation of polypeptide chains (Rules approved 1974)“. In: *Pure Appl. Chem.* 40.3 (1974), S. 291–308. DOI: [10.1351/pac197440030291](https://doi.org/10.1351/pac197440030291). URL: <http://www.chem.qmul.ac.uk/iupac/misc/ppep1.html> (zitiert auf S. 42).
- [Jcg08] J. C. F. G. I. M. Jcgm. „Evaluation of measurement data — Guide to the expression of uncertainty in measurement“. In: *Int. Organ. Stand. Geneva ISBN 50.* September (2008), S. 134. DOI: [10.1373/clinchem.2003.030528](https://doi.org/10.1373/clinchem.2003.030528). URL: <http://www.bipm.org/en/publications/guides/gum.html> (zitiert auf S. 8, 20, 31, 45).
- [JG78] G. H. Joblove, D. Greenberg. „Color Spaces for Computer Graphics“. In: *ACM siggraph Comput. Graph.* 12.3 (1978), S. 20–25. DOI: [10.1017/CBO9781107415324.004](https://doi.org/10.1017/CBO9781107415324.004) (zitiert auf S. 55).

- [JS03] C. R. Johnson, A. R. Sanderson. „A next step: Visualizing errors and uncertainty“. In: *IEEE Comput. Graph. Appl.* 23.5 (2003), S. 6–10. DOI: [10.1109/MCG.2003.1231171](https://doi.org/10.1109/MCG.2003.1231171) (zitiert auf S. 17).
- [KBE08] M. Krone, K. Bidmon, T. Ertl. „GPU-based Visualisation of Protein Secondary Structure.“ In: *Tpcg* (2008), S. 1–8. DOI: [10.2312/LocalChapterEvents/TPCG/TPCG08/115-122](https://doi.org/10.2312/LocalChapterEvents/TPCG/TPCG08/115-122) (zitiert auf S. 58).
- [KJB+17] L. Kocincová, M. Jarešová, J. Byška, J. Parulek, H. Hauser, B. Kozlíková. „Comparative visualization of protein secondary structures.“ In: *BMC Bioinformatics* 18.2 (2017), S. 23. DOI: [10.1186/s12859-016-1449-z](https://doi.org/10.1186/s12859-016-1449-z) (zitiert auf S. 12, 13).
- [Krz09] M. et al Krzywinski. „Circos: an Information Aesthetic for Comparative Genomics“. In: *Genome Res* 19.604 (2009), S. 1639–1645. DOI: [10.1101/gr.092759.109.19](https://doi.org/10.1101/gr.092759.109.19) (zitiert auf S. 50).
- [KS83] W. Kabsch, C. Sander. „Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features“. In: *Biopolymers* 22.12 (1983), S. 2577–2637. DOI: [10.1002/bip.360221211](https://doi.org/10.1002/bip.360221211) (zitiert auf S. 2, 7, 30, 38, 40).
- [LB00] M. Leitner, B. P. Buttenfield. „Guidelines for the display of attribute certainty“. In: *Cartogr. Geogr. Inf. Sci.* 27.1 (2000), S. 3–14. DOI: [10.1559/152304000783548037](https://doi.org/10.1559/152304000783548037) (zitiert auf S. 49).
- [LCPM97] G. Labesse, N. Colloc'h, J. Pothier, J. P. Mornon. „P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins.“ In: *Comput. Appl. Biosci.* 13.3 (1997), S. 291–295. DOI: [10.1093/bioinformatics/13.3.291](https://doi.org/10.1093/bioinformatics/13.3.291) (zitiert auf S. 11).
- [LV02] C. H. Lee, A. Varshney. „Representing Thermal Vibrations and Uncertainty in Molecular Surfaces“. In: *January* (2002), S. 80–90. DOI: [10.1117/12.458813](https://doi.org/10.1117/12.458813) (zitiert auf S. 16, 17).
- [Mac92] A. M. MacEachren. „Visualizing uncertain information“. In: *Cart. Perspect.* 13.13 (1992), S. 10–19. DOI: [10.1.1.62.285](https://doi.org/10.1.1.62.285) (zitiert auf S. 14).
- [Mar13] V. Marx. „Data visualization: ambiguity as a fellow traveler“. In: *Nat. Methods* 10.7 (2013), S. 613–615. DOI: [10.1038/nmeth.2530](https://doi.org/10.1038/nmeth.2530) (zitiert auf S. 50).
- [MLM+05] J. Martin et al. „Protein secondary structure assignment revisited: a detailed analysis of different assignment methods“. In: *BMC Struct. Biol.* 5.1 (2005), S. 17. DOI: [10.1186/1472-6807-5-17](https://doi.org/10.1186/1472-6807-5-17) (zitiert auf S. 11).
- [MMHT92] A. L. Morris, M. W. MacArthur, E. G. Hutchinson, J. M. Thornton. „Stereochemical quality of protein structure coordinates“. In: *Proteins Struct. Funct. Bioinforma.* 12.4 (1992), S. 345–364. DOI: [10.1002/prot.340120407](https://doi.org/10.1002/prot.340120407) (zitiert auf S. 20).
- [MRH05] A. M. Maceachren, A. Robinson, S. Hopper. „Visualizing geospatial information uncertainty: What we know and what we need to know“. In: *Cart. Geogr. Inf. Sci.* 32.3 (2005), S. 139–160. DOI: [10.1559/1523040054738936](https://doi.org/10.1559/1523040054738936) (zitiert auf S. 14).

- [MRO+12] A. M. Maceachren, R. E. Roth, J. O'Brien, D. Swingley, M. Gahegan. „Visual Semiotics and Uncertainty Visualisation: An Empirical Study“. In: *IEEE Trans. Vis. Comput. Graph.* 18.12 (2012), S. 1–10. DOI: [10.1109/TVCG.2012.279](https://doi.org/10.1109/TVCG.2012.279) (zitiert auf S. 14, 48).
- [NS12] G. E. Newman, B. J. Scholl. „Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias“. In: *Psychon. Bull. Rev.* 19 (2012), S. 601–607. DOI: [10.3758/s13423-012-0247-5](https://doi.org/10.3758/s13423-012-0247-5) (zitiert auf S. 50).
- [Pan01] A. Pang. „Visualizing uncertainty in geo-spatial data“. In: *Proc. Work. Intersect. between Geospatial Inf. Inf. Technol.* (2001), S. 1–14. DOI: [10.1.1.20.3823](https://doi.org/10.1.1.20.3823) (zitiert auf S. 8).
- [PRJ12] K. Potter, P. Rosen, C. R. Johnson. „From quantification to visualization: A taxonomy of uncertainty visualization approaches“. In: *IFIP Adv. Inf. Commun. Technol.* 377 AICT (2012), S. 226–247. DOI: [10.1007/978-3-642-32677-6_15](https://doi.org/10.1007/978-3-642-32677-6_15) (zitiert auf S. 14, 48).
- [PWL97] A. T. Pang, C. M. Wittenbrink, S. K. Lodha. „Approaches to uncertainty visualization“. In: *Vis. Comput.* 13.8 (1997), S. 370–390. DOI: [10.1007/s003710050111](https://doi.org/10.1007/s003710050111) (zitiert auf S. 8, 9, 14, 48).
- [Ric81] J. S. Richardson. *The anatomy and taxonomy of protein structure*. Bd. 34. Academic Press, Inc., 1981, S. 167–339. ISBN: 0-12-034234-0. URL: <http://kinemage.biochem.duke.edu/teaching/anatax/index.html> (zitiert auf S. 21).
- [Riv07] M. Riveiro. „Evaluation of uncertainty visualization techniques for information fusion“. In: *2007 10th Int. Conf. Inf. Fusion* (2007), S. 1–8. DOI: [10.1109/ICIF.2007.4408049](https://doi.org/10.1109/ICIF.2007.4408049) (zitiert auf S. 17).
- [RJ99] P. Rheingans, S. Joshi. „Visualization of Molecules with Positional Uncertainty“. In: *Data Vis. '99, Proc. Jt. EUROGRAPHICS - IEEE TCVG Symp. Vis.* Springer-Verlag, 1999, S. 299–306. ISBN: 978-3-211-83344-5, 978-3-7091-6803-5 (zitiert auf S. 15, 16).
- [Roc14] L. F. O. Rocha. „Toward a better understanding of structural divergences in proteins using different secondary structure assignment methods“. In: *J. Mol. Struct.* 1063.1 (2014), S. 242–250. DOI: [10.1016/j.molstruc.2014.01.060](https://doi.org/10.1016/j.molstruc.2014.01.060) (zitiert auf S. 2, 12).
- [RSS94] B. Rost, C. Sander, R. Schneider. „Redefining the goals of protein secondary structure prediction.“ In: *J. Mol. Biol.* 235.1 (1994), S. 13–26. DOI: [S0022-2836\(05\)80007-5](https://doi.org/S0022-2836(05)80007-5)[pii] (zitiert auf S. 11).
- [SAMG14] A. Sarikaya, D. Albers¹, MitchellJ., M. Gleicher. „Visualizing Validation of Protein Surface Classifiers“. In: *Comput Graph Forum* 33.3 (2014), S. 171–180. DOI: [10.1111/cgf.12373](https://doi.org/10.1111/cgf.12373) (zitiert auf S. 13).
- [Sec02] A. Secord. „Weighted Voronoi stippling“. In: *Proc. Second Int. Symp. Non-photorealistic Animat. Render. - NPAR '02 1* (2002), S. 37. DOI: [10.1145/508535.508537](https://doi.org/10.1145/508535.508537) (zitiert auf S. 77).

- [SNG+16] C. Schulz, A. Nocaj, J. Goertler, O. Deussen, U. Brandes, D. Weiskopf. „Probabilistic Graph Layout for Uncertain Network Visualization“. In: *IEEE Trans. Vis. Comput. Graph.* 2626.c (2016), S. 1–1. DOI: [10.1109/TVCG.2016.2598919](https://doi.org/10.1109/TVCG.2016.2598919) (zitiert auf S. 14, 15).
- [Sta14] wwPDB annotation Staff. „wwPDB Processing Procedures and Policies Document“. In: January (2014). URL: <http://cdn.wwpdb.org/wwpdb/docs/documentation/annotation/wwPDB-A-2014Jan03.pdf> (zitiert auf S. 7, 19, 20).
- [SZB+09] J. Sanyal, S. Zhang, G. Bhattacharya, P. Amburn, R. J. Moorhead. „A user study to compare four uncertainty visualization methods for 1D and 2D datasets“. In: *IEEE Trans. Vis. Comput. Graph.* 15.6 (2009), S. 1209–1218. DOI: [10.1109/TVCG.2009.114](https://doi.org/10.1109/TVCG.2009.114) (zitiert auf S. 14, 49).
- [TA08] A. Telea, D. Auber. „Code flows: Visualizing structural evolution of source code“. In: *Comput. Graph. Forum* 27.3 (2008), S. 831–838. DOI: [10.1111/j.1467-8659.2008.01214.x](https://doi.org/10.1111/j.1467-8659.2008.01214.x) (zitiert auf S. 13).
- [TBB+15] W. G. Touw, C. Baakman, J. Black, T. A. H. Te Beek, E. Krieger, R. P. Joosten, G. Vriend. „A series of PDB-related databanks for everyday needs“. In: *Nucleic Acids Res.* 43.D1 (2015), S. D364–D368. DOI: [10.1093/nar/gku1028](https://doi.org/10.1093/nar/gku1028) (zitiert auf S. 7, 26).
- [THMG05] J. Thomson, E. Hetzler, A. Maceachren, M. Gahegan. „A Typology for Visualizing Uncertainty“. In: *Vis. Data Anal.* 5669.January (2005), S. 146–157. DOI: [10.1117/12.587254](https://doi.org/10.1117/12.587254) (zitiert auf S. 14).
- [VHK+13] C. Vehlow, J. Hasenauer, A. Kramer, A. Raue, S. Hug, J. Timmer, N. Radde, F. J. Theis, D. Weiskopf. „iVUN: interactive Visualization of Uncertain biochemical reaction Networks“. In: *BMC Bioinformatics* 14 Suppl 1.Suppl 19 (2013), S2. DOI: [10.1186/1471-2105-14-S19-S2](https://doi.org/10.1186/1471-2105-14-S19-S2) (zitiert auf S. 16, 17).
- [WMDJ08] A. Wlodawer, W. Minor, Z. Dauter, M. Jaskolski. *Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures*. 2008. DOI: [10.1111/j.1742-4658.2007.06178.x](https://doi.org/10.1111/j.1742-4658.2007.06178.x) (zitiert auf S. 6, 21).
- [ZC06] T. Zuk, S. Carpendale. „Theoretical analysis of uncertainty visualizations“. In: 6060 (2006), S. 606007–606014. DOI: [10.1117/12.643631](https://doi.org/10.1117/12.643631) (zitiert auf S. 14, 48, 49, 55).
- [ZS15] Y. Zhang, C. Sagui. „Secondary structure assignment for conformationally irregular peptides: Comparison between DSSP, STRIDE and KAKSI“. In: *J. Mol. Graph. Model.* 55 (2015), S. 72–84. DOI: [10.1016/j.jmgm.2014.10.005](https://doi.org/10.1016/j.jmgm.2014.10.005) (zitiert auf S. 11, 41).

Alle URLs wurden zuletzt am 25. 03. 2017 geprüft.

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschrift (Matthias Braun)