Kapitel 4

Zusammenfassung und Ausblick

In dieser Arbeit werden verschiedene Methoden zur Generierung von Kontaktenergieparametern für die Proteinstrukturvorhersage verglichen. Überprüft werden folgende Fragestellungen:

- Ist eine gegebene Energiefunktion in der Lage, native von nicht-nativen Proteinstrukturen zu unterscheiden?
- Werden nicht-native Strukturen, die eine Ähnlichkeit zur nativen Struktur im Sinne eines hohen *Overlaps* aufweisen, als ähnlich erkannt?
- Wird eine gegebene native Struktur während einer Monte Carlo Simulation stabilisiert? Das heisst, die native Struktur sollte während einer solchen Simulation bei niedriger Temperatur den nativen Bereich nicht verlassen.
- Ist eine Energiefunktion in der Lage, zu einer gegebenen Sequenz eine Struktur mit nativen Eigenschaften in einer Monte Carlo Simulation aufzufinden?

Grundsätzlich wird unterschieden zwischen Proteinstrukturen, die zur Optimierung der Kontaktenergiefunktion verwendet wurden (gelernt) und solchen, die nicht zur Optimierung verwendet wurden. Ist eine Energiefunktion in der Lage, Strukturen, die nicht gelernt wurden, richtig zuzuordnen so ist sie *übertragbar*.
Wird eine Struktur mit niedrigster Energie für eine Zielsequenz gefunden ohne dass die experimentelle Proteinstruktur unter den alternativen Strukturen vorhanden war und wurde die Energiefunktion nicht auf diese Zielsequenz trainiert, so liegt eine echte Strukturvorhersage vor.
Folgende Methoden zur Berechnung der Kontaktenergieparameter werden verwendet:

- Eine Methode, die darauf abzielt den Boltzmann-gewichteten *Overlap Q* zwischen *Decoy* Strukturen und experimentellen Strukturen zu maximieren (boltz).

- Eine lineare Optimierung (LO), die auf der Lösung eines linearen Gleichungssystems basiert.

- Eine quasichemische Methode (QCM), die über die Häufigkeiten des Auftretens der möglichen Kontakte in nativen und nicht-nativen Proteinstrukturen die Kontaktenergieparameter bestimmt.

Von beiden Methoden werden verschiedene Varianten getestet. Um die Fähigkeit der Energiefunktionen native und nicht-native Proteinstrukturen richtig zuzuordnen, zu testen, werden verschiedene Sets an Proteinstrukturen verwendet.
Damit die Energiefunktionen in sinnvoller Weise trainiert und getestet werden können, ist es notwendig, dass die nicht-nativen Strukturen typische Proteinmerkmale aufweisen: der Abstand zwischen in der Sequenz benachbarten C_α Atomen sollte z.B. 3.8Å betragen. Sinnvoll ist auch, wenn typische Sekundärstrukturmerkmale wie α Helices und β Faltblätter vorliegen.
Eine einfache und effektive Methode solche nicht-nativen Strukturen zu erzeugen ist *Threading*. Hierbei wird eine Sequenz zu der Strukturen erzeugt werden sollen (die Zielsequenz) und die Struktur eines nativen Proteins zu einem Sequenz/Struktur-Paar (*Decoy*) vereinigt. Mit Hilfe von mehreren Zielsequenzen und einer größeren Zahl an nativen Proteinstrukturen lässt sich ein Set mit einer großen Anzahl von Sequenz/Struk-tur-Paaren erzeugen.
Am erfolgreichsten bei der richtigen Zuordnung dieser Paare als nativ oder nicht-nativ ist eine quasichemische Methode, welche berücksichtigt, dass zu den verschiedenen Zielsequenzen normalerweise unterschiedlich viele Decoys vorliegen (QCMw) (siehe Gleichung 2.35). Bei Verwendung des größten Proteinsets (Set_{1014}, dieses Set enthält 1014 Ketten von 965 verschiedenen Proteinen. Alle 1014 Ketten werden zur Erzeugung von *Decoys* verwendet, alle 202 einzelkettigen Proteine mit einer Länge kleiner gleich 200 Aminosäuren werden als Zielsequenzen verwendet), werden 70% der Zielsequenzen richtig als nativ erkannt. Diese Methode ist auch bezüglich der Übertragbarkeit den anderen Methoden überlegen: es genügt ein Training mit einem sehr kleinen Set an Strukturen um auch bei sehr viel größeren Sets gute Erkennung zu erreichen. Desweiteren genügt es, von allen *Decoys* nur die unähnlichsten für das Training zu verwenden. Für die untersuchten Sets an Strukturen können über 90% der Decoys beim Lernen ausgeschlossen werden. Die erreichte Erkennung (unter Verwendung aller Strukturen) bleibt hierbei konstant bzw. verbessert sich sogar. Werden bei Verwendung von Set_{1014} alle Strukturen mit einem *Overlap* $q > 0.2$ beim Lernen der Energieparameter ausgeschlossen (dies entspricht 94% der Strukturen), so liegt die Erkennung immer noch bei 70%.

In der einfachsten Form besteht die Kontaktenergiefunktion aus je einem Energieparameter für jedes der 210 möglichen Aminosäurepaare. Erweiterungen sind problemlos möglich. Wird z.B. zwischen verschiedenen Abständen der Residuen entlang der Sequenz unterschieden, also für ein Aminosäurepaar der Residuen i und j der Kontaktenergieparameter in Abhängigkeit von $|j-i|$ gewählt, so lässt sich die Erkennung bei geeigneter Wahl von Sequenzabstandsbereichen deutlich verbessern. Für Set_{135} lässt sich die Erkennung von 70% bei Verwendung von nur einem Bereich auf 85% bei Verwendung von zwei Abstandsbereichen steigern. Bei Set_{420} verbessert sich hierbei die Erkennung von 52% auf 65%.

Eine wichtige Eigenschaft der Kontaktenergiefunktionen ist die Fähigkeit Strukturen, die Ähnlichkeit zur nativen Struktur im Sinne eines hohen *Overlaps* aufweisen, als ähnlich zu erkennen. Um diese Eigenschaft zu testen werden Zielsequenzen für die *Decoys* mit hoher Ähnlichkeit vorliegen näher untersucht: weist der *Decoy* höchster Ähnlichkeit die niedrigste Energie aller *Decoys* zu dieser Sequenz auf, so wurde dieser *Decoy* erfolgreich als ähnlich erkannt. Beim Test der 14 Sequenzen mit den ähnlichsten *Decoys* (*overlap* $q \geq 0.7$) zeigt sich, dass bei sieben Sequenzen bei keiner der untersuchten Methoden eine Vorhersage mit einem *Overlap* größer 0.5 vorliegt. Bei Verwendung von QCMw Set_{45} wird nur bei drei Sequenzen der ähnlichste *Decoy* als solcher erkannt. Dies ist ein typisches Problem von Kontaktenergiefunktionen: Native Proteine werden gut als solche erkannt, bei geringen Abweichungen von der nativen Struktur jedoch versagt die Energiefunktion bei der Aufgabe die Ähnlichkeit zu erkennen.
Eine Funktion, die Proteinstrukturen hinsichtlich der Ähnlichkeit zur nativen Struktur sinnvoll bewertet, sollte sich auch für Monte Carlo Simulationen eignen. Wird für eine solche Simulation eine mittels der linearen Optimierung und Set_{420} (welches Crambin nicht enthält) trainierte Energiefunktion verwendet und dient die native Struktur von Crambin als Startpunkt, so ergibt sich eine Struktur mit einem *Overlap* von $q = 0.78$, und einer C_α cRMSD von 4.44Å. Dient als Startpunkt der Monte Carlo Simulation eine Struktur ohne native Eigenschaften, so liegt eine echte Strukturvorhersage vor. Unter Verwendung der Energiefunktion QCMw Set_{45} wird für Crambin ein *Overlap* von $q = 0.56$ bei einer cRMSD von 6.66Å erreicht. Die Struktur liegt somit zwar nicht im „nativen Bereich“, zeigt jedoch schon deutliche Ähnlichkeit zur nativen Struktur. Auffälligerweise werden die besten Ergebnisse erzielt, wenn kleine Proteinsets zum Lernen der Energiefunktion verwendet werden. Hier zeigt sich also der gleiche Trend wie bei der Erkennung der nativen Strukturen: ein kleines repräsentatives Set enthält bereits sehr viel Information bezüglich der allgemeinen Struktur-Sequenz Abhängig-

keit in Proteinen. Für das Protein 2erl (*Mating Pheromone Er-1*) ergibt sich (bei Verwendung von QCMw Set_{45}) ein *Overlap* von $q = 0.62$ bei einer cRMSD von 5.84Å. Für 1orc (*Cro Repressor Insertion Mutant K56-[Dgevk]*) liegt mit einem *Overlap* von $q = 0.40$ und einer cRMSD von 9.56Å keine sinnvolle Vorhersage vor.
Monte Carlo Simulationen können auch dafür genutzt werden, um Strukturen verschiedenster Ähnlichkeit zu einer gegebenen Sequenz zu erzeugen. So können z.B. alle Strukturen einer Monte Carlo Trajektorie als *Decoys* zum Training einer Kontaktenergiefunktion verwendet werden. Wird die Simulation z.B. mit einer nativen Struktur gestartet, so lässt sich die Ähnlichkeit der erzeugten Strukturen über die verwendete Energiefunktion und die Simulationstemperatur steuern. Faltungssimulationen mit Energiefunktionen, die mit solchen Strukturen trainiert wurden, erzeugen keine besseren Strukturen als wenn das Training der Energiefunktion mit *Threading* Strukturen erfolgt. Diese Tatsache lässt vermuten, dass die Auswahl der *Decoys* zum Training der Energiefunktion nicht den limitierenden Faktor darstellt.

Ein wichtiger Punkt beim Vergleich von Proteinstrukturen ist das verwendete Distanz- bzw. Ähnlichkeitskriterium. Verschiedene Distanzkriterien werden hinsichtlich Ihrer Fähigkeit C_α-C_α Abstandsverteilungen von Proteinen wiederzugeben überprüft. Es zeigt sich, dass die *power distance*, welche die Atomabstände innerhalb der Proteinstrukturen berücksichtigt, eine sehr gute Beschreibung der Proteinstrukturen liefert. Um dieses Distanzkriterium anwenden zu können müssen jedoch die Atomabstände bekannt sein. Die Kontaktdistanz D_{cont}, welche über die Beziehung $D_{\text{cont}} = 1 - q$ mit dem *Overlap* q in Beziehung steht, erreicht zwar nicht die Qualität der *power distance* bei der Beschreibung von Proteinstrukturen, ist dafür aber sehr schnell anwendbar. Diese Eigenschaft macht sie besonders geeignet für *Threading*, da hier eine sehr große Anzahl an Proteinstrukturen erzeugt und verglichen wird.
Das Prinzip der Kontaktenergiefunktion lässt sich auf vielerlei Weise erweitern. Neben der hier verwendeten Erweiterung der Energiefunktion auf mehrere Sequenzabstandsbereiche ist auch eine Erweiterung auf mehrere Abstandsbereiche im Raum möglich. Statt also z.B. „Kontakt“ zu definieren, wenn der C_α-C_α Abstand 11Å nicht überschreitet, kann z.B. unterschieden werden zwischen einem Abstand kleiner $R1_c$, einem Abstand zwischen $R1_c$ und $R2_c$ und einem Abstand größer $R2_c$ (kein Kontakt). Auch können verschiedene Abstandsbereiche für verschiedene Aminosäurepaare verwendet werden.