Kapitel 3

Ergebnisse

3.1 Eigenschaften von *Decoys* und nativen Proteinen

Möchte man Strukturvorhersagen machen, so ist ein Wissen über Unterschiede zwischen nativen und nicht-nativen Strukturen von großem Nutzen. Wird eine wissensbasierte Energiefunktion verwendet und die Energiefunktion mit nativen und nicht-nativen Strukturen trainiert, so ist es wichtig, dass sich die nicht-nativen Strukturen nicht trivial von den nativen Strukturen unterscheiden. Die C_α-C_α Abstände von in der Peptidkette benachbarten Aminosäuren z.B. betragen 3.8Å. Kommen in einer nicht-nativen Struktur andere Werte für diese C_α-C_α Abstände vor, so kann die Struktur von vornherein als nicht-nativ eingeordnet werden. Wird die Energiefunktion mit nicht-nativen Strukturen mit solchen falschen C_α-C_α Abständen trainiert, so ist sie vielleicht in der Lage die trainierten Strukturen richtig zu unterscheiden. Es ist jedoch fraglich, ob nicht-native Strukturen mit korrekten C_α-C_α Abständen von in der Kette benachbarten Residuen ebenfalls als nicht-nativ erkannt werden. Ein sehr wichtiger Punkt beim Erzeugen der *Decoys* ist also, dass diese physikalisch sinnvoll und zu nativen Proteinstrukturen ähnlich sind.

3.1.1 Vergleich nativer Strukturen mit *Decoy* Strukturen aus der *Threading* Methode

Werden Strukturen mittels *Threading* erzeugt, so erhält man eine große Zahl an physikalisch sinnvollen *Decoys*. Typische Proteinmerkmale, wie die Sekundärstrukturen sind enthalten. Ein Nachteil ist jedoch, dass hierbei nur eine geringe Zahl an Strukturen mit hoher Ähnlichkeit zu der nativen Struktur einer vorgegebenen Zielsequenz erzeugt wird.

Abb. 3.1 zeigt ein Histogramm der jeweils maximalen *Overlaps* $q_{\max}$ aller *Decoys* für die 202 Zielsequenzen aus dem Set_{1014}. Für ein gutes Trainingsset sind nicht nur hohe maximale *Overlaps* wichtig. Gleichzeitig müssen ähnliche *Decoys* auch in ausreichender Zahl vorhanden sein. Die zweite Kurve in Abb. 3.1 zeigt das Histogramm für die Verteilung der q-Werte aller *Decoys*. Insgesamt handelt es sich um rund $24.9 \cdot 10^6$ *Decoys*. Wie man den Kurven entnehmen kann, liegen nur für wenige Zielsequenzen *Decoys* mit hoher Ähnlichkeit vor. Von allen *Decoys* beträgt der höchste *Overlap* $q = 0.993$, die hierzu korrespondierende cRMSD der C_α Atome beträgt 0.14Å. Wie

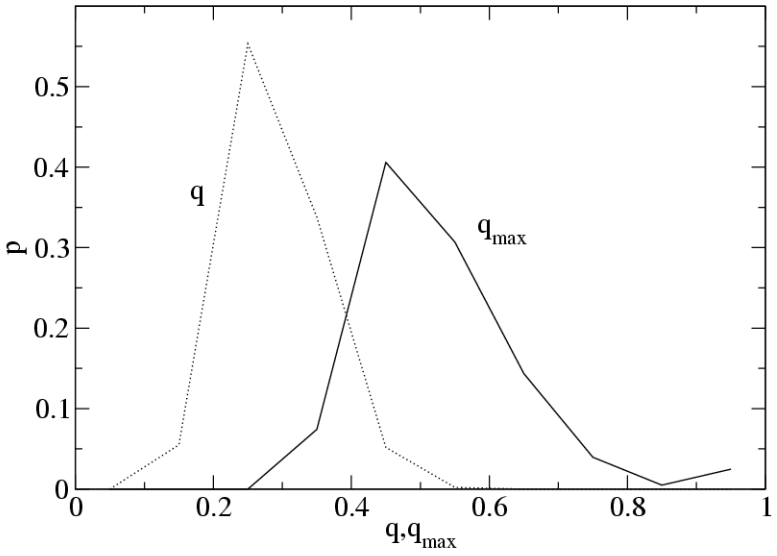

Abbildung 3.1: Wahrscheinlichkeitsverteilung der *Overlaps* q aller *Decoys* sowie der *Decoys* mit größtem *Overlap* (q_{max}) zur nativen Struktur für das Set_{1014}. Es gilt das C_α Kontaktkriterium, mit einem Abstandskriterium von $r_c = 11$Å. Als in Kontakt gelten nur Residuen mit einem Sequenzabstand von mindestens dis_{seq}=3.

bereits erwähnt, ist es wichtig, dass sich native und nicht-native Strukturen nicht trivial voneinander unterscheiden. Sind zum Beispiel alle *Decoys* in ihrer Struktur offener als die nativen Proteine, so ist es ein leichtes diese zu erkennen. Für diesen Fall würde es genügen nur Energieparameter mit negativem Vorzeichen zu wählen. Die kompakteste Struktur ist dann automatisch die energieärmste. Auch der umgekehrte Fall (alle *Decoys* sind kompakter als die nativen Strukturen) wäre denkbar. Wird für die Erzeugung eines *Decoys* aus einer großen Struktur ein kleiner Bereich herausgeschnitten, so ist zu erwarten, dass dieser Bereich eher eine offene Konformation aufweist, als ein kleines vollständiges Protein. Für eine Sequenz in Set_{1014} (2erl) haben acht der erzeugten *Decoys* im Mittel sogar weniger als zwei Kontakte pro Residuum. Die große Zahl an *Decoys* mit eher offener Struktur spiegelt sich in Abb. 3.2 wieder. Aufgetragen ist das

Histogramm der mittleren Anzahl der Kontakte pro Aminosäure für alle *Decoys* und für alle nativen Zielsequenzen. Wie Abb. 3.3 zu entnehmen ist, liegt meistens jedoch zu jeder Sequenz noch eine recht große Zahl an kompakteren Strukturen vor.

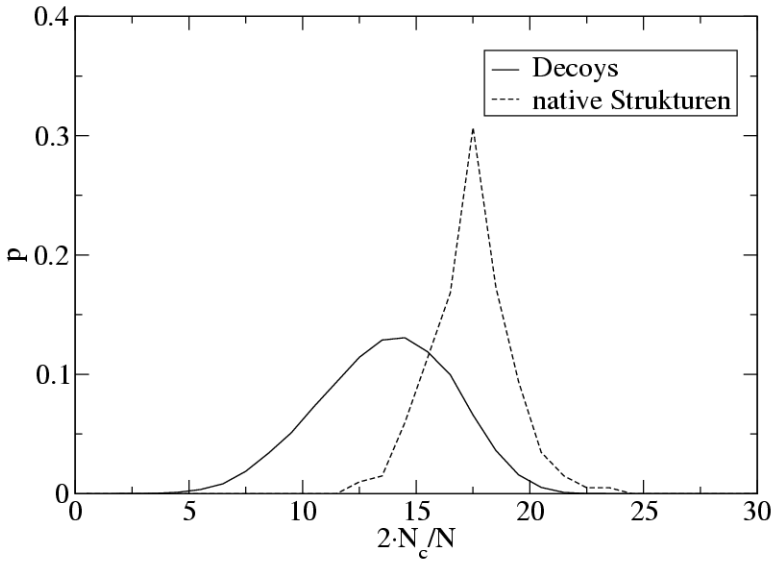

Abbildung 3.2: Histogramm der mittleren Anzahl der Kontakte für alle nativen Zielsequenzen aus Set_{1014} (gestrichelte Linie) sowie für alle mit diesem Set erzeugten *Decoys* (durchgezogene Linie).

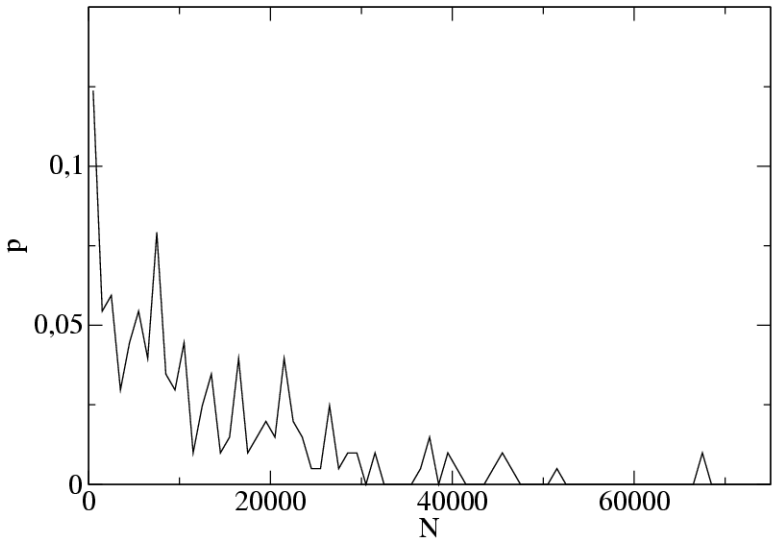

Abbildung 3.3: Histogramm der Anzahl der *Decoys N*, die eine kompaktere Struktur als die native Struktur besitzen, für die 202 einzelnen Zielsequenzen aus Set_{1014}.

3.1.2 Packungsdichte in nativen Proteinen

Die mittlere Anzahl der Kontakte einer Struktur hängt von der Form und der Länge des Proteins ab. Eine sehr kompakte Struktur hat viele Kontakte, eine offene Struktur weniger. Die meisten Proteine sind eher kompakt und das Verhältnis von Oberfläche zu Volumen verringert sich mit zunehmender Sequenzlänge. In Abbildung 3.4 ist die mittlere Zahl der Nachbarn der einzelnen Aminosäuren in Abhängigkeit von der Kettenlänge für Kristallstrukturen aufgetragen. Verwendet wurden alle 202 einzelkettigen Proteine mit einer Länge $N \leq 200$ aus Set_{1014}. Wie zu erwarten, nimmt die mittlere Anzahl der Kontakte mit der Länge zunächst zu, um dann auf einem relativ konstanten Niveau zu bleiben. Trägt man für die Kristallstrukturen die mittleren Kontaktzahlen aus Abb. 3.4(a) gegen $\mathrm{N}^{-1/3}$ auf, so lassen sich die Werte relativ gut mit einer linearen Funktion anpassen:

$$a - bN^{-\frac{1}{3}} \tag{3.1}$$

Hierbei ist $a = 25.9$ und $b = 42.3$, der Korrelationskoeffizient beträgt $r = -0.60$. Abb. 3.4(b) zeigt die mittlere Zahl der Nachbarn für die 135 einzelkettigen Proteine mit einer Länge $N \leq 200$ aus dem $\mathrm{Set}_{\mathrm{NMR}}$. Die Parameter der angepassten Funktion 3.1 sind hier $a = 24.5$ und $b = 39.1$. Diese Werte unterscheiden sich nicht sehr von denen für Kristallstrukturen. Die Streuung der Werte ist jedoch für NMR Strukturen höher. Hier ist der Korrelationskoeffizient mit $r = -0.46$ deutlich geringer als bei den Kristallstrukturen.
Werden alternative Strukturen mittels *Threading*, unter Verwendung des C_α Kontaktmodells erzeugt, so weisen die alternativen Strukturen zwar eher eine offene Struktur auf (siehe 3.1.1), die Packungsdichten bewegen sich jedoch ungefähr im gleichen Bereich wie bei nativen Proteinstrukturen.
Werden Strukturen mit der Monte Carlo Methode erzeugt, so ist über das abstoßende Potential gewährleistet, dass einzelne Residuen sich nicht zu Nahe kommen. Somit ist auch hier die Anzahl der Kontakte begrenzt, wobei der Wert generell etwas zu hoch liegt. Um in Monte Carlo Simulationen die Kontaktzahl nicht zu stark ansteigen zu lassen, enthält die Energiefunktion hier einen zusätzlichen Term, der hohe Kontaktzahlen mit einer positiven Energie belegt (siehe Gleichung 2.10).

3.2 Proteinmodelle

Das Modell einer Proteinstruktur sollte die für eine gegebene Anwendung relevanten Eigenschaften der Struktur möglichst genau wiedergeben. Verschiedene Protein-

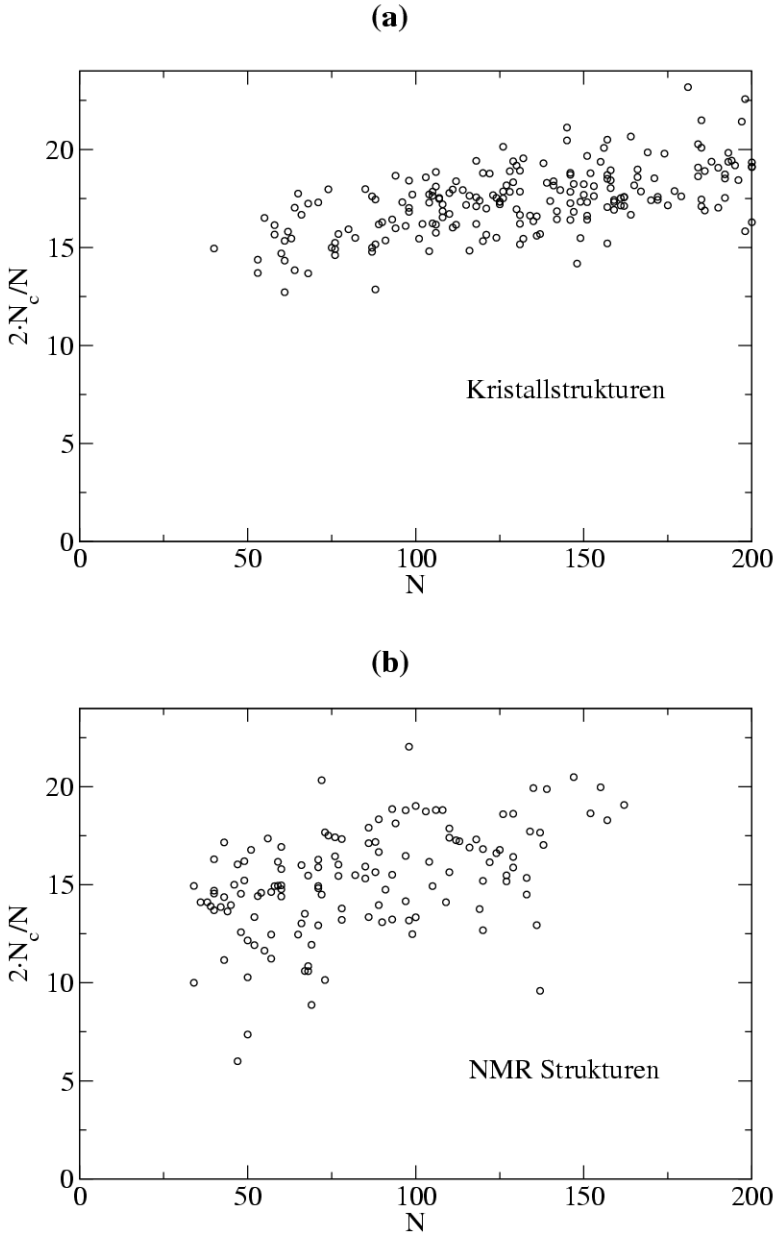

Abbildung 3.4: Mittlere Anzahl der Kontakte $2 \cdot \frac{N_c}{N}$ in Abhängigkeit von der Kettenlänge N für Kristallstrukturen (Set_{1014}) bzw. NMR-Strukturen ($\mathrm{Set}_{\mathrm{NMR}}$) unter Verwendung des C_α Kontaktkriteriums.

modelle werden hier getestet.

Ein diskretes Proteinmodell Das in 2.4.1 vorgestellte Proteinmodell beinhaltet starke Vereinfachungen: die Konformation einer Kette wird nur durch die Winkel α und τ bestimmt. Der Winkel α ist hierbei der Pseudo-Bindungswinkel zwischen drei aufeinanderfolgenden C_α Atomen, der Winkel τ bezeichnet den Pseudo-Torsionswinkel zwischen vier aufeinanderfolgenden C_α Atomen. Um eine der nativen Struktur möglichst ähnliche Struktur im Raume der α/τ Winkelpaare zu generieren wird der *Build up* Algorithmus von Park und Levitt verwendet. Ein Satz an α/τ Winkelpaaren (hier sechs) dient dem sukzessiven Aufbau der Kette an C_α Atomen. In jedem Schritt wird ein C_α Atom an die wachsende Kette unter Verwendung aller Winkelpaare angefügt. Bei sechs verwendeten Winkelpaare also in sechs unterschiedlichen Konformationen. Um alle denkbaren Konformationen zu erzeugen, müssten nun an jede erzeugte Kette wieder das nächste C_α Atom unter Verwendung aller Winkelpaare angefügt werden. Da dies sehr schnell zu einer extrem hohen Anzahl an Konformationen führt, wird nach jedem neu hinzugefügten C_α Atom nur eine maximale Anzahl N_{keep} an Konformationen weiterverwendet. Um den *Build up* Algorithmus möglichst sinnvoll verwenden zu können müssen die verwendeten Winkelpaare sorgfältig gewählt werden. Hier werden die Winkelpaare mittels einer Monte Carlo Methode optimiert (siehe 2.4.2.1). Diese Optimierung erfolgt mit $N_{\text{keep}} = 200$, beim Aufbau der Peptidkette werden also maximal 200 Strukturen je Schritt gespeichert.
Sowohl bei der cRMSD als auch bei der Kontaktdistanz D_{cont} können unphysikalisch kleine Atomabstände auftreten. Bei der *power distance* haben kleine Abstände ein hohes Gewicht, wodurch das Auftreten von solchen Zusammenstößen zwischen Atompaaren sehr wirkungsvoll verhindert wird.
Um Zusammenstöße zu verhindern wird hier für die cRMSD und die Kontaktdistanz D_{cont} ein abstoßendes Potential eingeführt. Tritt während des *Build up* Algorithmus eine Struktur mit einem C_α-C_α Atomabstand kleiner als ein Grenzwert r_c auf, so wird die Struktur verworfen. Der niedrigste C_α-C_α Abstand bei den 774 verwendeten Kristallstrukturen (siehe 2.4.2.2) liegt bei 2.61Å, als Grenzwert wird $r_c = 2.6$Å verwendet. Im folgenden werden Modelle mit und ohne ein solches abstoßendes Potential verglichen. Es werden somit fünf Optimierungen der (α, τ) Winkel durchgeführt: ohne abstoßendes Potential für die cRMSD, die Kontaktdistanz D_{cont} und die *power distance* D_{pow} sowie mit abstoßendem Potential für die cRMSD und D_{cont}.

3.2.1 Proteinmodelle ohne abstoßendes Potential

In Tabelle 3.1 sind die für die verschiedenen Distanzmaße optimierten Winkel für die Modelle ohne abstoßendes Potential aufgeführt. Mit der cRMSD als Distanzmaß lassen sich (α,τ) Winkel finden, die die Modellstrukturen mit einer durchschnittlichen cRMSD von 1.57Å an die Kristallstrukturen anpassen. Nur bei zwei Proteinen ist der Wert größer als 2.0Å, wobei der höchste Wert 2.1Å beträgt.
Mit der Kontaktdistanz D_{cont} werden die Modellstrukturen im Mittel mit $D_{\text{cont}} = 0.23$ und einem Maximum von $D_{\text{cont}} = 0.46$ angepasst. Bei der *Power distance* beträgt der Mittelwert 0.19, der schlechteste 0.33.

	cRMSD	D_{cont}	D_{pow}
(α,τ) Winkel	71.5, 71.2	83.3, 72.4	85.5, −62.4
	87.9, 55.6	92.0, 39.8	94.9, 96.0
	104.2, −111.0	115.0, −163.1	103.6, 163.0
	104.6, 36.6	118.0, −64.6	115.8, −152.2
	124.0, −160.0	128.5, 111.5	119.6, −22.0
	129.5, 128.8	129.7, −119.2	125.2, 126.8
mittlere Distanz	1.57 Å	0.23	0.19

Tabelle 3.1: Optimierte (α,τ) Winkel für ein Modell ohne abstoßendes Potential.

Wie gut sich eine Modellstruktur an die Kristallstruktur mit dem *Build up* Algorithmus anpassen lässt, hängt neben den (α,τ) Winkeln davon ab, wie viele Strukturen maximal bei jedem Schritt gespeichert werden, also welcher Wert N_{keep} verwendet wird (siehe 2.4.2). Abb. 3.5 zeigt die Abhängigkeit der durchschnittlichen Distanz der Modellstrukturen von N_{keep} am Beispiel der cRMSD. Verwendet wurde $\text{Set}_{\text{disk,keep}}$. Hierbei handelt es sich um 38 Proteine aus Set_{45} (siehe Anhang). Bis zu einem Wert von $N_{\text{keep}} = 200$ nimmt die cRMSD schnell ab. Bei $N_{\text{keep}} = 10$ beträgt die mittlere cRMSD 1.75Å, bei $N_{\text{keep}} = 200$ beträgt sie 1.56Å. Wird N_{keep} auf 1000 erhöht, beträgt die mittlere cRMSD noch 1.54Å. Die Verwendung von $N_{\text{keep}} = 200$ scheint also sinnvoll, da mit weiterer Erhöhung von N_{keep} nur noch eine geringe Verbesserung der cRMSD erreicht wird.
Die Anpassung der Modellproteine an ihre Kristallstrukturen in Bezug auf die verwendeten Distanzmaße funktioniert also relativ gut. Die Frage ist, wie gut andere strukturelle Eigenschaften von Kristallstrukturen wiedergegeben werden. Abb. 3.6 zeigt die C_α-C_α Abstandsverteilungen für die Modellstrukturen der 774 Sequenzen, angepasst unter Verwendung der verschiedenen Distanzkriterien sowie für die Kristallstrukturen.

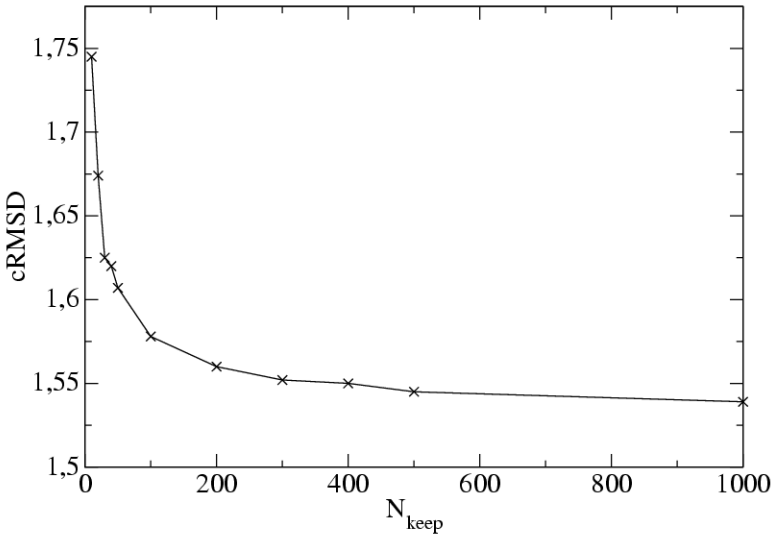

Abbildung 3.5: Abhängigkeit der erreichten cRMSD in Abhängigkeit von N_{keep} (maximal gespeicherte Anzahl der Strukturen pro Schritt des *Build up* Algorithmus (siehe 2.4.2)).

Es sind nur C_α-C_α Paare mit einem Sequenzabstand von mindestens $\text{dis}_{\text{seq}} = 3$ berücksichtigt. Aufgetragen ist die normalisierte Häufigkeit der C_α-C_α Abstände gegen r:

$$f(r) = \frac{N(r,\Delta r)}{r^2 \Delta r} \tag{3.2}$$

Hierbei ist $N(r,\Delta r)$ der Anteil an Paaren mit einem C_α-C_α Abstand im Intervall $[r, r+\Delta r]$, für Δr wird ein Wert von 0.1 Å verwendet. Es zeigt sich, dass die Verteilung der realen Strukturen und die der Modelle große Unterschiede aufweisen. Kleine C_α-C_α Abstände unter 2.6Å kommen bei den realen Strukturen aus sterischen Gründen nicht vor. In Abb. 3.6 sind Residuen, die innerhalb der Peptidkette benachbart bzw. nur durch ein Residuum getrennt sind, von der Betrachtung ausgeschlossen. Für diesen Fall beträgt der minimale C_α-C_α Abstand 2.96Å. Die sterische Hinderung wird von der *power distance*, welche Abweichungen von kleinen Abständen hoch wichtet, sehr gut wiedergegeben. Die cRMSD gibt diese Eigenschaft weniger gut wieder. Abweichungen werden unabhängig vom Abstand gewichtet. Bei der Kontaktdistanz D_{cont} führt eine hohe Zahl an gemeinsamen Kontakten zu einem niedrigen Wert. Dies ist vermutlich der Grund dafür, dass die Verteilung bei kleinen Abständen von diesem Kriterium am schlechtesten wiedergegeben wird.
Werden nur große Abstände über 13Å betrachtet, so ergibt sich ein anderes Bild. Die Funktion 3.2 für die Kristallstrukturen lässt sich für $r > 13$ Å mit einer Exponential-

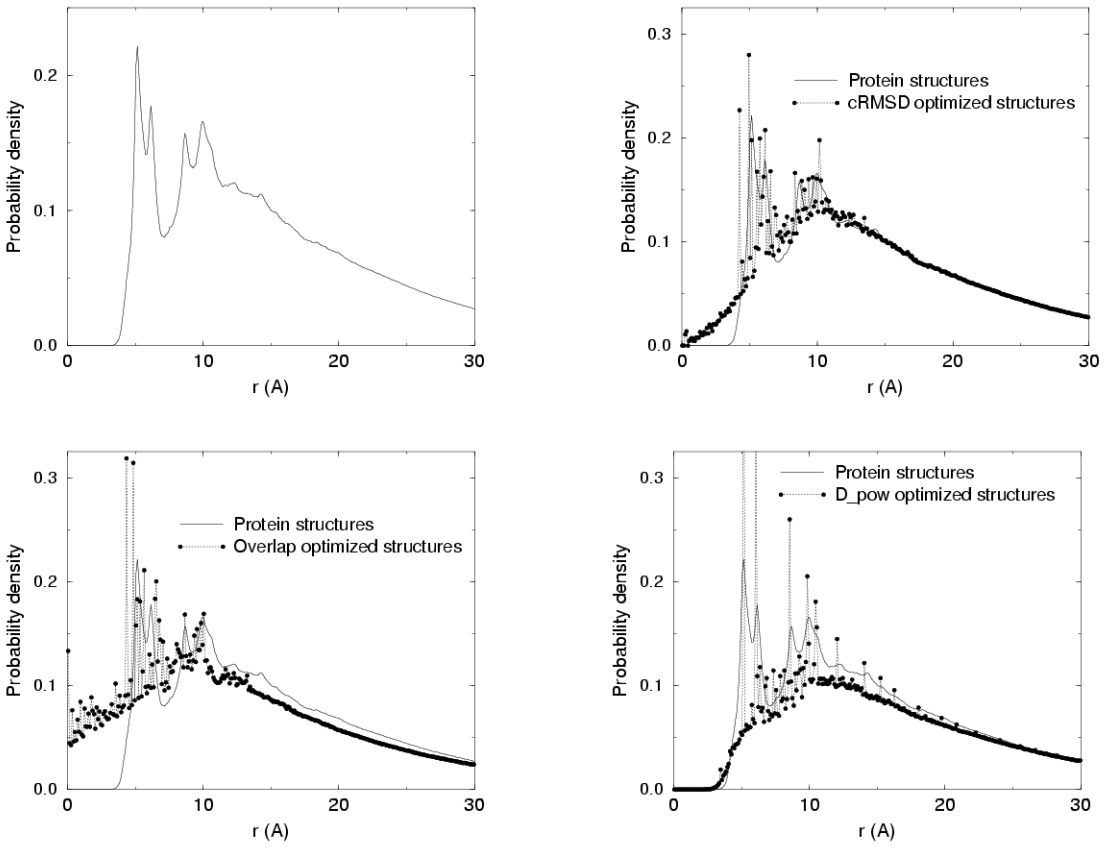

Abbildung 3.6: C_α-C_α Abstandsverteilungen realer Proteinstrukturen und der mittels verschiedener Ähnlichkeitsmaße angepassten Modellstrukturen. Aufgetragen ist die normalisierte Häufigkeit der C_α-C_α Abstände gegen r: $f(r) = \frac{N(r,\Delta r)}{r^2 \Delta r}$. Bei der *power distance* liegt der höchste Peak ausserhalb der Grafik. Dieser hat für $r = 5.0$ einen Wert von $f(r) = 0.53$.

funktion anpassen:

$$f(r) \propto \exp(-r/\xi) \tag{3.3}$$

Die optimale Anpassung erfolgt mit $\xi = 8.30 \pm 0.01$. Der Mittelwert aller C_α-C_α Abstände beträgt 28.8 Å. Für die mittels der cRMSD angepassten Strukturen ergibt sich ebenfalls $\xi = 8.30 \pm 0.01$ Å, der Mittelwert der C_α-C_α Abstand beträgt 28.7Å. Die Werte von Kristall- und Modellstrukturen stimmen hier sehr gut überein. Große Abstände lassen sich mit der cRMSD also gut wiedergeben. Für die mittels *power distance* angepassten Strukturen hingegen erfolgt die beste Anpassung mit $\xi = 8.50 \pm 0.01$ Å, der mittlere C_α-C_α Abstand beträgt 29.8 Å. Diese Werte sind relativ hoch im Vergleich zu denen der Kristallstrukturen was die Tatsache widerspiegelt, dass große Abstände geringer gewichtet und somit schlechter angepasst werden. Für die mittels D_{cont} angepassten Strukturen ergeben sich sogar noch höhere Werte. Die Anpassung an die Expo-

nentialfunktion ergibt $\xi = 10.69 \pm 0.01$ Å, der mittlere C_α-C_α Abstand beträgt 32.8Å. Ein Grund für diese hohen Werte liegt vermutlich in der Definition des *Overlaps*: die Zahl der gemeinsamen Kontakte wird durch die maximale Zahl der Kontakte dividiert. Ist die Zahl der gemeinsamen Kontakte hoch, so kann der *Overlap* durch Minimierung der maximalen Zahl der Kontakte weiter vergrößert werden, was einen niedrigen Wert für die Kontaktdistanz D_{cont} ergibt. Eine geringe Zahl der Kontakte wirkt sich dann positiv auf D_{cont} aus. Die Feinstruktur der Abstandsverteilung wird von keinem der drei Kriterien gut wiedergegeben. Bei den realen Strukturen liegt bei 5.2-6.2Å ein Bereich bevorzugter Interaktion vor, welcher durch einen Doppelpeak gekennzeichnet ist. Bei 7.5Å liegt ein Minimum vor. Residuen, die mit dem betrachteten Residuum in Kontakt stehen, bewirken hier vermutlich eine sterische Hinderung. Die Tatsache, dass sowohl bei der cRMSD als auch bei D_{cont} unphysikalisch niedrige Abstände vorkommen legt nahe, ein abstoßendes Potential für geringe Abstände einzuführen.

3.2.2 Proteinmodelle mit abstoßendem Potential

Es ergeben sich folgende optimierte Winkel für das Modell mit abstoßendem Potential:

	cRMSD	D_{cont}
(α,τ) Winkel	83.7, 62.0	82.7, 63.9
	95.9, 41.6	104.8, 169.4
	109.7, −151.9	106.9, 25.9
	110.8, −104.4	113.3, −137.3
	129.3, 186.7	113.8, −70.6
	134.0, 120.9	128.0, 109.9
mittlere Distanz	1.54 Å	0.23

Tabelle 3.2: Optimierte (α,τ) Winkel für ein Modell mit abstoßendem Potential.

Die erreichte mittlere Kontaktdistanz D_{cont} der Modellstrukturen entspricht dem des Modells ohne Abstoßung. Die mittlere cRMSD ist sogar ein wenig besser. Der Raum an möglichen Konformationen verkleinert sich natürlich beim Übergang zum Modell mit abstoßendem Potential. Daher kann die erreichte Ähnlichkeit prinzipiell nicht besser werden. Die kleinere cRMSD ist auf eine bessere Optimierung der Winkel zurückzuführen. Die C_α-C_α Abstandsverteilungen (siehe Abb. 3.7) unterscheiden sich ebenfalls kaum vom Modell ohne Abstoßung, außer dass sich im Bereich $r < 2.6$Å keine C_α-C_α Paare finden, da Strukturen mit solchen Abständen im *Build up* Algorithmus abgelehnt werden.

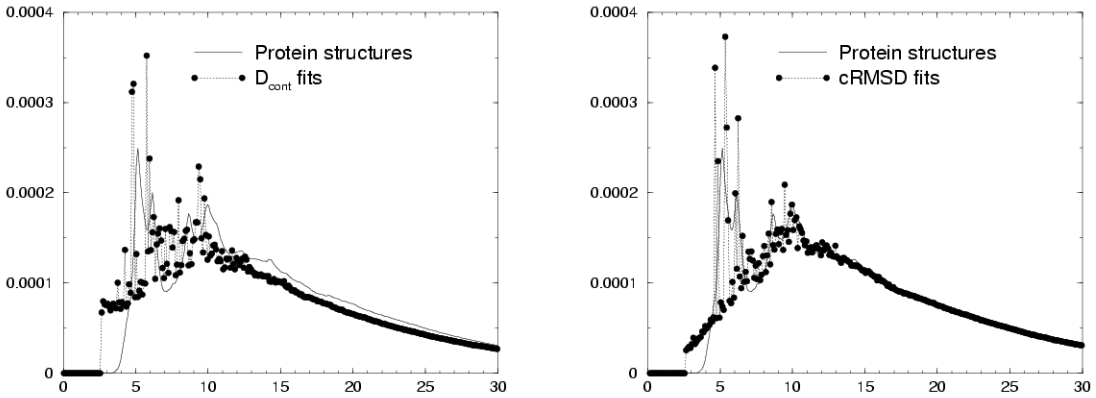

Abbildung 3.7: C_α-C_α Abstandsverteilungen realer Proteinstrukturen und der mittels verschiedener Ähnlichkeitsmaße angepassten Modellstrukturen. Es wird ein abstoßendes Potential verwendet: Während der *Build up* Prozedur werden alle Strukturen mit einem C_α-C_α Abstand von $r < 2.6$Å abgelehnt. Aufgetragen ist die normalisierte Häufigkeit der C_α-C_α Abstände gegen r: $f(r) = \frac{N(r,\Delta r)}{r^2 \Delta r}$.

Eine Anpassung des Bereiches $r > 13$Å an eine Exponentialfunktion ergibt $\xi = 8.29 \pm 0.01$ Å für die cRMSD und $\xi = 10.62 \pm 0.01$ Å für die Kontaktdistanz D_{cont}. Die entsprechenden mittleren C_α-C_α Abstände betragen 28.8Å bzw. 32.5Å. Das Modell wird durch die Einführung des abstoßenden Potentials, abgesehen vom Fehlen der kleinen C_α-C_α Abstände, nicht signifikant beeinflusst.

3.3 Eigenschaften der Energiefunktionen

Eine wichtige Eigenschaft der Energiefunktionen ist die Fähigkeit native von nicht-nativen Proteinstrukturen zu unterscheiden. Hierauf wird in 3.3.1 eingegangen. Diese Eigenschaft alleine reicht aber weder aus um ähnliche Strukturen zu erkennen noch um sinnvolle Faltungssimulationen durchzuführen. Hierfür wird eine korrelierte Energielandschaft benötigt, also eine Energielandschaft mit einer Form, die möglichst der eines Trichters entspricht. Wie die Energiefunktion bei der Erkennung von Strukturen mit nativen Eigenschaften funktioniert wird in 3.3.5 dargelegt. In 3.4.1 wird anschließend gezeigt, wie sich die Energiefunktion auf Faltungssimulationen anwenden lässt.

3.3.1 Erkennung von nativen Proteinstrukturen

Als erstes wird getestet, ob die Energiefunktion in der Lage ist native Proteinstrukturen von nicht-nativen Strukturen zu unterscheiden. Im einfachsten Fall erfolgt das Training und der Test der Energiefunktion mit dem selben Satz an Strukturen. Mit steigender Zahl der nativen Strukturen und der zugehörigen *Decoys* wird es immer schwieriger die Energiefunktion in sinnvoller Weise zu trainieren, da zunehmend mehr Bedingungen gleichzeitig erfüllt werden müssen. Für diesen Test wird für das Training und den Test jeweils das gleiche Set aus den vier verschiedenen Proteinsets verwendet (siehe 2.7).
Eine sinnvolle Energiefunktion sollte übertragbar sein, also auch solche Paare an nativen und nicht-nativen Strukturen unterscheiden können, die nicht trainiert wurden. Um die Übertragbarkeit der Energiefunktionen zu überprüfen wird getestet, ob native Proteinstrukturen, die nicht trainiert wurden, als nativ erkannt werden bzw. ob bei einer Hinzunahme von *Decoys*, die nicht trainiert wurden, die native Struktur weiterhin erkannt wird. Je mehr native Strukturen und zugehörige *Decoys* gelernt werden und je repräsentativer diese Strukturen sind, desto einfacher sollte es im allgemeinen sein ungelernte Strukturen richtig zuzuordnen. Um die Übertragbarkeit zu testen wird für das Training ein gegebenes Protein-Set verwendet und die Energiefunktion dann durch Anwendung auf ein anderes, größeres Set getestet. Von den vier Sets an Kristallstrukturen Set_{45}, Set_{135}, Set_{420} und Set_{1014} stellt jedes Set eine Untermenge der größeren Sets dar. Wird die Energiefunktion also mit einem größeren Set getestet, so sind die trainierten Strukturen alle auch im Testset enthalten.
Auch ein Test mit einem kleineren Set kann aufschlussreich sein. In diesem Fall sollte die Erkennung im allgemeinen schlechter sein als wenn mit dem kleineren Testset direkt trainiert wurde, da Bedingungen gelernt werden, die im Test nicht enthalten sind.

3.3.1.1 Erzeugen der Energieparameter durch eine Boltzmann-gewichtete Optimierung

Werden die Energieparameter mittels der Boltzmann-gewichteten Optimierung (siehe 2.9) erzeugt, so ergeben sich für die verschiedenen Proteinsets folgende Erkennungen:

	Trainingsset			
Testset	Set_{45}	Set_{135}	Set_{420}	Set_{1014}
Set_{45}	98%	93%	93%	89%
Set_{135}	79%	95%	93%	89%
Set_{420}	61%	68%	92%	87%
Set_{1014}	51%	57%	80%	86%

Tabelle 3.3: Erkennung von nativen Proteinen unter Verwendung der Boltzmann-gewichteten Optimierung (siehe 2.9). Es gilt das C_α Kontaktkriterium.

Wird das größte Set an Proteinen verwendet, so werden immerhin noch 86% aller Zielsequenzen erkannt.

3.3.1.2 Erzeugen der Kontaktenergieparameter mittels einer linearen Optimierung

Bei der Erzeugung der Parameter mit Hilfe des linearen Gleichungssystems wird die Energie einer Struktur mit deren Ähnlichkeit zur nativen Konformation korreliert (siehe 2.10). Ein entscheidender Punkt ist die Verwendung einer geeigneten Funktion $f(q)$ für diese Korrelation. Mit den einfachen Beziehungen $f(q) = 1 - q$ bzw. $f(q) = -q$ werden nur äusserst geringe Erkennungen der nativen Proteine erreicht. Tabelle 3.4 zeigt die Erkennungen für die verschiedenen Proteinsets unter Verwendung des C_α Kontaktmodells. Die Energiefunktion wird auf ein gegebenes Set trainiert und die Energieparameter mit dem selben Set getestet.
Unter Verwendung einer Exponentialfunktion $-De^{(\beta \cdot q)}$ wird die Erkennung erheblich verbessert. Abb. 3.8 zeigt die Erkennung in Abhängigkeit vom Exponenten β für Set_{135}, Set_{420} und Set_{1014}. Der Faktor D beeinflusst nur die Skalierung der Parameter und ist hier auf $D = 1$ gesetzt. Für Set_{135} ist die maximale Erkennung 73.2%. Diese wird mit $\beta = 12$ sowie mit $\beta = 15$ bis $\beta = 20$ erreicht. Für Set_{420} wird mit $\beta = 15$, $\beta = 16$ und $\beta = 17$ eine maximale Erkennung von 57.0% erreicht. Mit $\beta = 18$, $\beta = 19$ und $\beta = 20$ ist die Erkennung mit 56.5% geringfügig geringer, wobei aber nur jeweils

	Für Training und Test verwendetes Set			
$f(q)$	Set_{45}	Set_{135}	Set_{420}	Set_{1014}
$1-q$	18%	0%	0%	0%
$-q$	13%	1%	1%	1%

Tabelle 3.4: Erkennung der nativen Proteinstrukturen unter Verwendung von $f(q) = 1-q$ und $f(q) = -q$. Das Training und der Test der Energiefunktionen erfolgt jeweils mit dem selben Proteinset. Als Kontaktkriterium wird das C_α Modell verwendet.

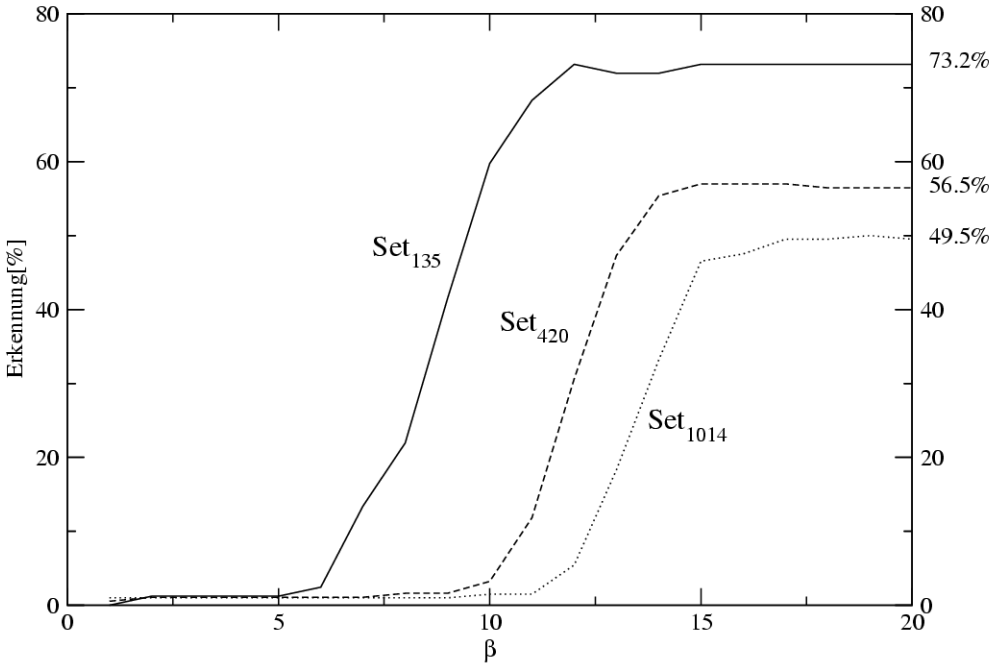

Abbildung 3.8: Erkennung unter Verwendung einer Exponentialfunktion $-e^{(\beta \cdot q)}$ für die Optimierung der Kontaktenergieparameter nach Gleichung 2.24 in Abhängigkeit von Parameter β. Es gilt das C_α Kontaktkriterium.

eine native Proteinstruktur weniger erkannt wird. Für Set_{1014} wird mit $\beta = 19$ eine maximale Erkennung von 50.0% erreicht. Mit $\beta = 17$, $\beta = 18$ und $\beta = 20$ ergibt sich eine Erkennung von 49.5%.
Abb. 3.9 zeigt entsprechende Exponentialfunktionen im Bereich $q \in [0,1]$. Zur besseren Vergleichbarkeit wird $D = e^{-\beta}$ verwendet. Die Funktionen sind dann so skaliert, dass $f(1) = -1$. Die Erkennung erreicht ihr Optimum im betrachteten Bereich bei

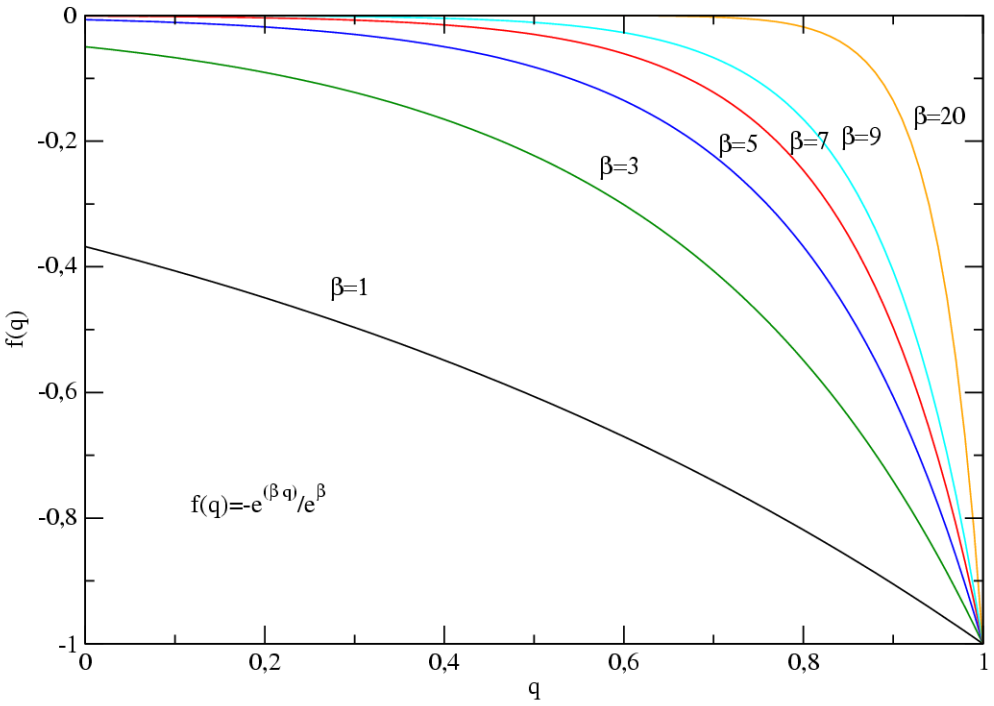

Abbildung 3.9: Schätzfunktion für die Energie von Proteinstrukturen in Abhängigkeit vom *Overlap* q, $f(q) = -De^{(\beta \cdot q)}$ für verschiedene Werte β. Zur besseren Vergleichbarkeit ist $D = e^{-\beta}$ verwendet. Die Funktionen sind dann so skaliert, dass $f(1) = -1$.

hohen Werten für β. Für die Optimierung der Kontaktenergieparameter wird *Threading* unter Verwendung eines der drei Proteinsets Set_{135}, Set_{420} oder Set_{1014} aus Abschnitt 2.7 verwendet. Kontakt wird über das C_α Kontaktkriterium definiert (siehe 2.2.1). Die *Overlaps* q der so erzeugten *Decoys* liegen fast ausschließlich im Bereich $q \in [0.1, 0.6]$ (siehe 2.4). In diesem Bereich ist die Exponentialfunktion $f(q) = -e^{(\beta \cdot q)}$ im Verhältnis zu $f(1)$ sehr klein und annähernd konstant. Für $\beta = 20$ ist das Verhältnis $\frac{f(0.6)}{f(1)} = 3.4 \cdot 10^{-4}$.
Es liegt also die Vermutung nahe, dass eine vergleichbare Erkennung mit einer Stufenfunktion erreicht werden kann:

$$f(q) = \begin{cases} 0 & : \text{für } q < 1 \\ -1 & : \text{für } q = 1 \end{cases} \tag{3.4}$$

In Tabelle 3.5 sind die Erkennungen für verschiedene Kombinationen aus Lern- und Testset aufgeführt. Wird für das Lernen der Energiefunktion und den Test auf Erkennen der nativen Strukturen das gleiche Set verwendet, so ist für Set_{135} die Erkennung mit 73.2% identisch zu der höchsten Erkennung unter Verwendung der Exponentialfunktion. Für Set_{420} ergibt sich eine Erkennung von 56.5%. Dieser Wert ist geringfügig kleiner als die beste Erkennung unter Verwendung der Exponentialfunktion (57.0%). Mit 49.0% für Set_{1014} erreicht die Stufenfunktion auch hier annähernd die Erkennung von 50% mit der Exponentialfunktion.
Die Übertragbarkeit, also die Fähigkeit der Energiefunktion Strukturen, die nicht gelernt wurden richtig als nativ oder nicht-nativ einzuordnen, ist relativ gut.
Es reicht, die Energiefunktion mit Set_{45} zu trainieren. Mit den auf diese Weise bestimmten Kontaktenergieparametern ist die Erkennung bei Anwendung auf die vier betrachteten Sets mindestens ebenso gut, als wenn die Energieparameter für die jeweils vollständigen Sets gelernt werden. Werden die Parameter z.B. auf Set_{1014} angewandt, so ist die Erkennung 51%. Werden die Energieparameter mittels des kompletten Set_{1014} gelernt, so ist die Erkennung nur 49%.

	Trainingsset			
Testset	Set_{45}	Set_{135}	Set_{420}	Set_{1014}
Set_{45}	91%	73%	80%	82%
Set_{135}	73%	73%	68%	70%
Set_{420}	58%	55%	57%	57%
Set_{1014}	51%	47%	50%	49%

Tabelle 3.5: Erkennung von nativen Proteinen unter Verwendung der Stufenfunktion 3.4 für die Optimierung der Kontaktenergieparameter nach Gleichung 2.24. Es gilt das C_α Kontaktkriterium.

Verwendung eines Polynoms in $(1-q)$ Werden die Parameter mit Hilfe des linearen Gleichungssystems unter Verwendung eines Polynoms in $(1-q)$ (siehe 2.25) erzeugt, so ist die Wahl des Polynomgrades Δ von großer Bedeutung. Je höher der Grad des Polynoms, desto besser lässt sich der *Overlap* q mit der Energie korrelieren, d.h desto kleiner sollte die Differenz $\left|\left|\mathbf{A}^t\vec{u} - \vec{f}(\vec{q})\right|\right|$ sein. Wird der Grad Δ jedoch zu hoch

gewählt, so treten numerische Probleme auf.
Hier wird das C_α- mit dem *all atom* Kontaktkriterium verglichen (siehe 2.2.1).

C_α Kontaktmodell Tabelle 3.6 zeigt die Erkennung von Zielsequenzen in Abhängigkeit vom Grad Δ des Polynoms für das C_α Modell. Der Unterschied zwischen rechter und linker Seite des Gleichungssystems für verschiedene Grade des Polynoms (siehe Gleichung 2.25) unter Verwendung der verschiedenen Protein Sets ist in Tabelle 3.7 angegeben. Es zeigt sich, dass für das Set_{135} der Unterschied bis zu einem Grad von acht abnimmt. Von $\Delta = 8$ auf $\Delta = 9$ nimmt der Unterschied zu. Dies deckt sich mit der Tatsache, dass bei einem Grad von neun die Konditionszahl der Matrix $\mathbf{A^2}$ größer ist als die reziproke Maschinengenauigkeit von 10^{12}. $\mathbf{A^2}$ ist also schlecht konditioniert (siehe 2.10.2). Ein Polynom bis zum Grade 9 scheint aber auch völlig auszureichen.
Abbildung 3.10 zeigt verschiedene Polynome zusammen mit dem Histogramm der *Overlaps* der verwendeten Strukturen. Für die Optimierung relevant ist nur der Bereich von q, in dem eine größere Zahl an *Decoys* auftritt. Der Häufigkeitsverteilung von q ist zu entnehmen, dass dieser Bereich sich ungefähr von $q = 0.1$ bis $q = 0.5$ erstreckt. Ab einem Grad von $\Delta = 4$, ändert sich die Funktion in diesem Bereich nur noch geringfügig. Auch zeigt sich, dass die Differenz zwischen rechter und linker Seite des Gleichungssystems ab einem Grad von $\Delta = 4$ für dieses Set nur noch geringfügig absinkt. Die Erkennung steigt ab hier auch nicht weiter an.
Werden die anderen Proteinsets verwendet, so steigt die Erkennung nicht kontinuierlich an. Für Set_{420} ist ein Polynom mit $\Delta = 5$ im angegebenen Bereich optimal. Beim Set_{1014} wird die maximale Erkennung von 43% schon bei $\Delta = 6$ erreicht. Dass über einen Grad von 11 hinaus die Erkennung sich noch stark verbessert ist nicht zu erwarten.

Δ	1	2	3	4	5	6	7	8	9	10	11
Set_{135}	4%	0%	44%	70%	70%	70%	70%	70%			
Set_{420}	2%	0%	2%	51%	54%	51%	52%	52%	53%	52%	53%
Set_{1014}	2%	0%	1%	27%	33%	43%	36%	42%	34%	35%	43%

Tabelle 3.6: Erkennung von nativen Proteinen unter Verwendung der linearen Optimierung in Abhängigkeit vom verwendeten Grad Δ des Polynoms (siehe Gleichung 2.25). Es gilt das C_α Kontaktkriterium.

Werden Energieparameter, die mit einem kleinen Set erzeugt wurden, auf ein größeres Set angewendet, so ist die Erkennung deutlich schlechter, als wenn für das Training

	$\|\mathbf{A}^t\vec{u} - \vec{f}(\vec{q})\|$			
Δ	1	2	3	4
Set_{135}	0.0701278561	0.01418010805	0.009330461246	0.00916185208
Set_{420}	0.07162577339	0.01128757348	0.004596485088	0.004228259449
Set_{1014}	0.07335409073	0.01128213032	0.003443783293	0.002906352048
Δ	5	6	7	8
Set_{135}	0.00915624347	0.00915605962	0.009156037409	0.009156025287
Set_{420}	0.004205646594	0.004199537472	0.00419301191	0.004189582819
Set_{1014}	0.002872140604	0.002866274127	0.002860445318	0.002855848966
Δ	9	10	11	
Set_{135}	0.009156027732	0.009156025183	0.009156027245	
Set_{420}	0.004188560263	0.004187744373	0.004187386822	
Set_{1014}	0.002854383621	0.002853314151	0.002852809253	

Tabelle 3.7: Mit zunehmendem Grad Δ des Polynoms (siehe Gleichung 2.25) nimmt die Differenz zwischen rechter und linker Seite des Gleichungssystems 2.21 theoretisch ab. Da jedoch mit endlicher Genauigkeit gearbeitet wird, kann beim Berechnen mit dem Computer auch eine Vergrößerung eintreten. Angegeben ist die Differenz des Gleichungssystems in Abhängigkeit vom Grad Δ des Polynoms unter Verwendung des C_α Kontaktkriteriums für verschiedene Proteinsets.

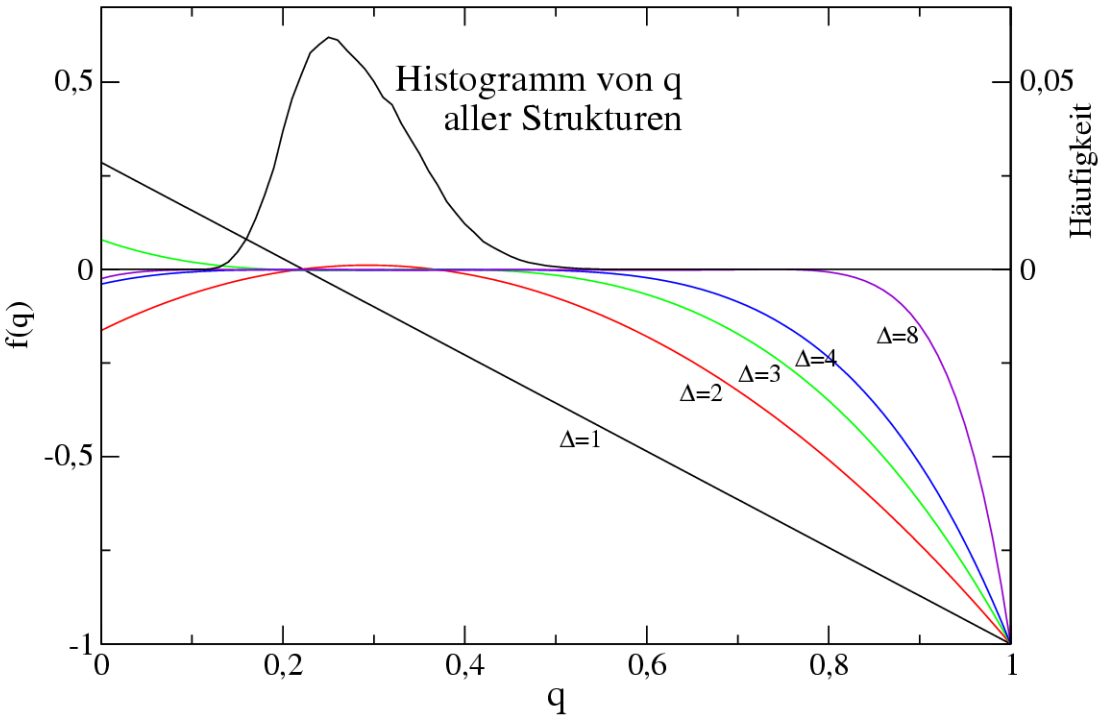

Abbildung 3.10: Polynome in $(1-q)$ mit verschiedenem Grad Δ (siehe Gleichung 2.25) für Set_{135} unter Verwendung des C_α Kontaktkriteriums sowie das Histogramm der *Overlaps* aller Strukturen die mit diesem Set erzeugt werden.

das komplette Testset verwendet wird. Für die im Trainingsset gelernten Sequenzen, werden im größeren Testset mehr *Decoys* erzeugt. Es müssen also *Decoys* aussortiert werden, die nicht gelernt wurden. Desweiteren sind im Testset Sequenzen enthalten, die überhaupt nicht gelernt wurden. Folgende Erkennungen ergeben sich für die verschiedenen Kombinationen aus Trainings- und Testset:

	Trainingsset			
Testset	Set_{45}	Set_{135}	Set_{420}	Set_{1014}
Set_{45}	82%	76%	76%	73%
Set_{135}	52%	70%	68%	60%
Set_{420}	36%	51%	52%	46%
Set_{1014}	29%	42%	45%	42%

Tabelle 3.8: Zum Testen einer gegebenen Energiefunktion wird diese mit einem Set an Proteinen (dem Lernset) trainiert und einem (anderen oder auch dem gleichen) Set an Proteinen getestet (dem Testset). Hier werden verschiedene Kombinationen aus Trainings- und Testset zum Testen von Energiefunktionen verwendet. Die Energieparameter sind mit Hilfe der linearen Optimierung (siehe 2.10) unter Verwendung des C_α Kontaktkriteriums berechnet.

Das Polynom zum Lernen der Energiefunktion enthält für Set_{45} sechs, für die anderen Sets acht Koeffizienten. Abb. 3.11 zeigt die 210 Energieparameter. Für Parameter mit besonders hohen Beträgen sind die zugehörigen Aminosäurepaare angegeben.

all atom **Kontaktmodell** Verwendet man das *all atom* Kontaktmodell, so steigt die Erkennung der nativen Strukturen stark an (siehe Tabelle 3.9). Wie in 2.2.4 erklärt,

Δ	1	2	3	4	5	6	7	8	9	10	11
Set_{135}	52%	23%	94%	98%	98%	98%	98%	98%			
Set_{420}	30%	5%	5%	92%	94%	94%	94%	94%	94%	94%	94%
Set_{1014}	30%	8%	2%	74%	94%	93%	94%	94%	93%	94%	93%

Tabelle 3.9: Erkennung von nativen Proteinen unter Verwendung der linearen Optimierung (siehe 2.24) in Abhängigkeit vom verwendeten Grad Δ des Polynoms (siehe 2.25) für das *all atom* Modell.

ist eine bessere Erkennung schon deshalb zu erwarten, da die *Decoys* bei Verwendung des *all atom* Modells eher von „natürlichen“ Proteinen abweichende Eigenschaften aufweisen, was von der Energiefunktion erkannt werden kann. Für die Erzeugung

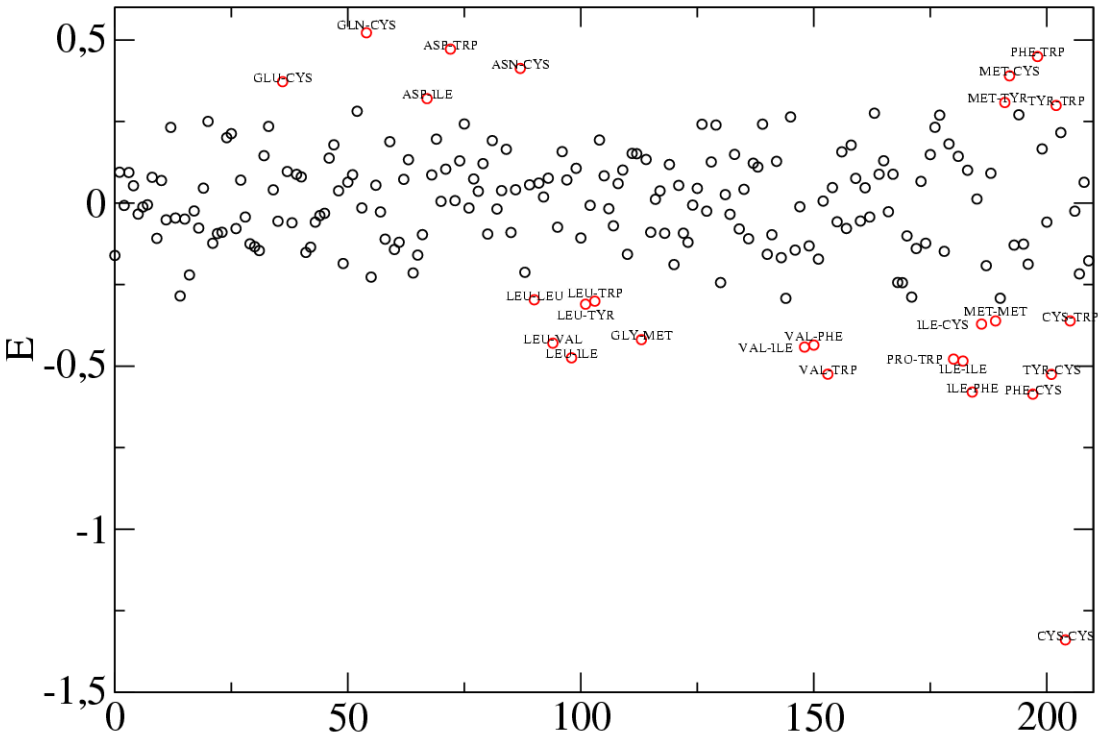

Abbildung 3.11: Die Energieparameter für Set_{135}.
Die Energieparameter sind mit Hilfe der linearen Optimierung (siehe 2.10) unter Verwendung eines Polynoms vom Grade $\Delta = 8$ (siehe Gleichung 2.25) berechnet. Es gilt das C_α Kontaktkriterium. Für Energieparameter mit besonders hohen Beträgen ist das zugehörige Aminosäurepaar angegeben

eines *Decoys* wird eine Sequenz mit einer Struktur zu einem Sequenz/Struktur Paar verknüpft. Bei Verwendung des C_α Modells bleibt die Struktur des Proteinrückgrats hierbei unverändert. Wird jedoch das *all atom* Modell verwendet, so wird in der Regel die Struktur des Proteinrückgrat verändert, wodurch wahrscheinlich eher energetisch ungünstige Strukturen entstehen. Je mehr die Sequenz der Zielsequenz von der Sequenz des Decoyproteins abweicht, desto größer ist im allgemeinen die Verzerrung der *Decoy* Struktur. Bei der nativen Struktur tritt dieses Problem nicht auf, was bedeutet dass diese schon durch die Tatsache, dass sie nicht verzerrt wurde leichter erkannt werden kann.

Unterschiedliche Wichtungen für verschiedene Proteinstrukturen Die Erkennung der nativen Proteinstruktur kann für unterschiedliche Sequenzen unterschiedlich schwierig sein. Wird eine native Struktur nicht erkannt, so kann durch Erhöhung der Gewichte aller zu einer Sequenz gehörenden Strukturen unter Umständen eine Erken-

nung erreicht werden (siehe 2.10.1). Werden z.B. die Energieparameter durch Minimieren der Abweichungen eines linearen Gleichungssystems (lineare Optimierung, siehe 2.10) erzeugt und die Gewichte nicht erkannter Sequenzen iterativ nach 2.10.1 erhöht, so lässt sich die Erkennung für das Set_{135} unter Verwendung des C_α Kontaktkriteriums von 70% auf 80% steigern. Abb. 3.13 zeigt die Wichtungsfaktoren w aller 82 Zielsequenzen aus Set_{135}, geordnet nach der Sequenzlänge. Das Gewicht von 32 Sequenzen wurde während der Rechnung nicht geändert. Da in jedem Schritt der Iteration das Gewicht der nicht erkannten Sequenzen erhöht wird, sinkt gleichzeitig das relative Gewicht der erkannten Sequenzen. Diese 32 Sequenzen scheinen also besonders unproblematisch bei der Erkennung zu sein. Bei neun Sequenzen wurde das Gewicht in jedem Schritt erhöht und keine Erkennung erreicht (siehe Tabelle 3.10 und Abb. 3.12).

1cuk	*E. coli* Ruva Protein (Helikase)
1lis	Lysin
1a1x	Mtcp-1
1hyp	*Hydrophobic Protein From Soybean*
1hoe	α-Amylase Inhibitor Hoe-467A.
1msi	*Antifreeze Glycoprotein Qae(HPLC 12)*
1orc	*Cro Repressor Insertion Mutant K56-[Dgevk]*
2igd	*Protein G IgG-Binding Domain*
1vie	*Dihydrofolate Reductase*

Tabelle 3.10: Werden die Gewichte nicht erkannter Sequenzen iterativ erhöht (siehe Gleichung 2.10.1), so steigt die Erkennung für Set_{135} von 70% auf 80% an. Die Gewichte von neun Sequenzen werden hierbei in jedem Schritt erhöht ohne dass eine Erkennung erreicht wird.

3.3.1.3 Erzeugen der Kontaktenergieparameter mittels einer quasichemischen Methode

Werden die Parameter mit der quasichemischen Methode (siehe 2.11) erzeugt, so ist die Erkennung besser als bei Verwendung der linearen Optimierung (siehe 2.10). Werden alle Strukturen gleich gewichtet, also Gleichung 2.33 verwendet, so ergeben sich die in Tabelle 3.11 aufgeführten Erkennungens. Auffällig ist, dass die Erkennung für das Set_{45} am schlechtesten ist, wenn für das Training ebenfalls Set_{45} verwendet wird. Erfolgt das Training mit einem der anderen Sets so steigt die Erkennung von 80% auf 89%. Werden also für das Lernen der Energiefunktion zusätzliche Strukturen verwen-

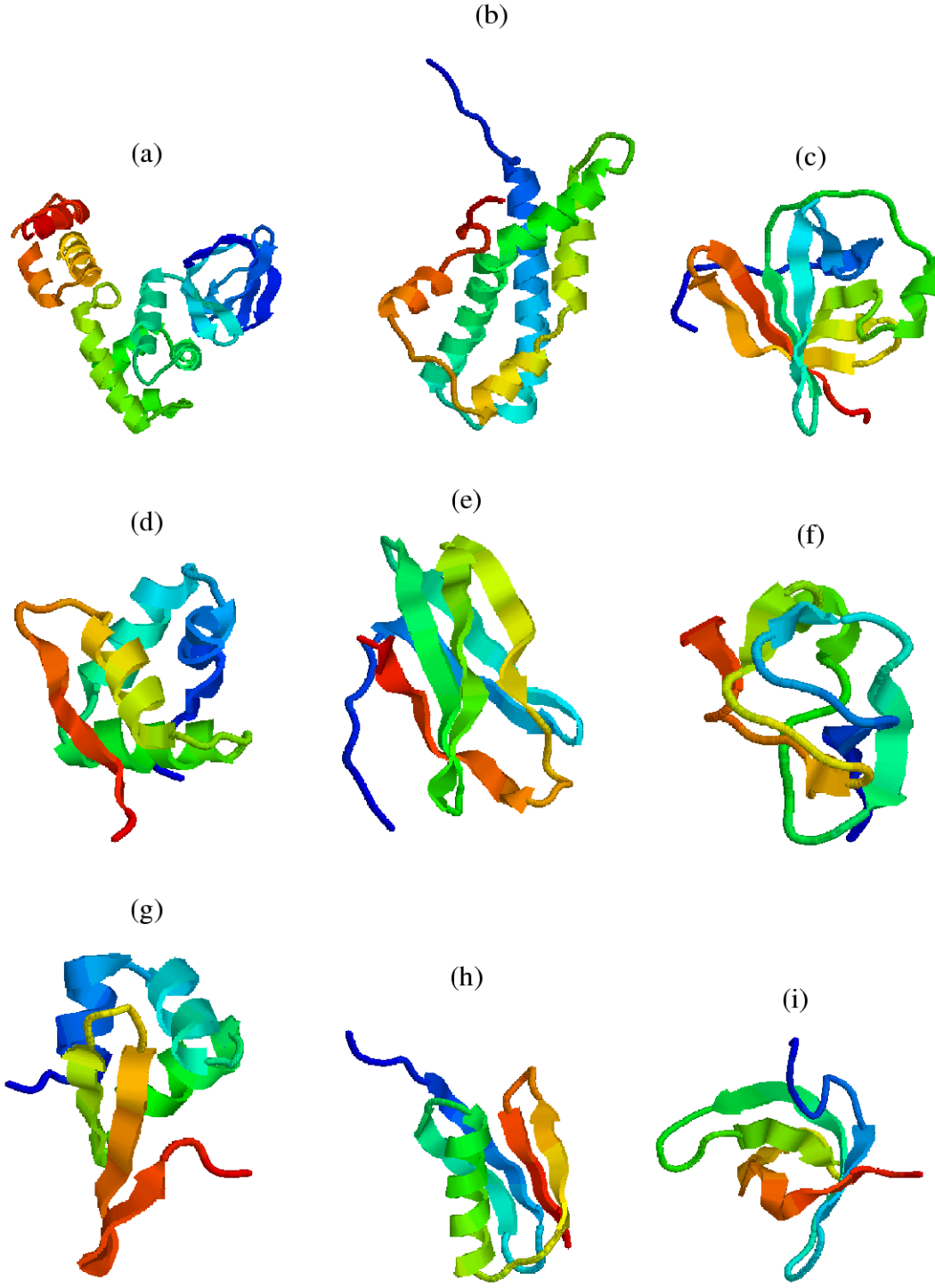

Abbildung 3.12: Werden die Gewichte nicht erkannter Sequenzen iterativ erhöht (siehe Gleichung 2.10.1), so steigt die Erkennung für Set_{135} von 70% auf 80% an. Die Gewichte von neun Sequenzen werden hierbei in jedem Schritt erhöht ohne dass eine Erkennung erreicht wird: (a): 1cuk, (b): 1lis, (c): 1a1x, (d): 1hyp, (e): 1hoe, (f): 1msi, (g): 1orc, (h): 2igd, (i): 1vie

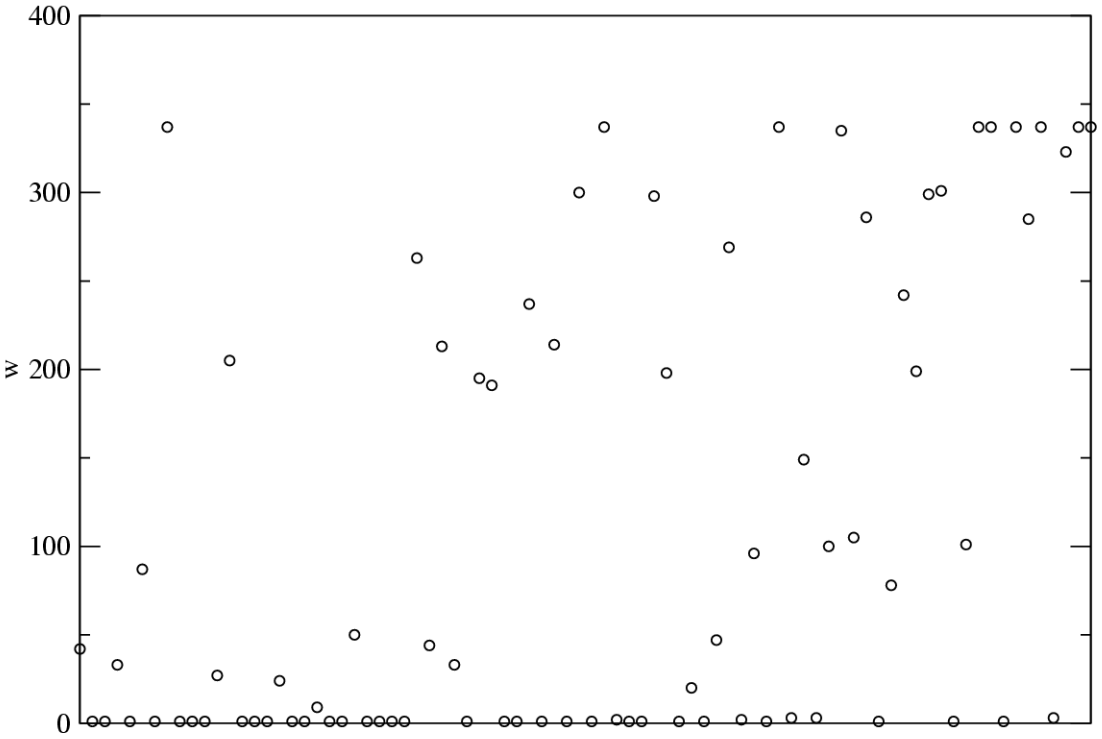

Abbildung 3.13: Die Erkennung der nativen Proteine lässt sich durch das Einführen unterschiedlicher Wichtungsfaktoren für die verschiedenen Sequenzen verbessern (siehe Gleichung 2.10.1). Für das Set_{135} steigt bei Verwendung der linearen Optimierung die Erkennung von 70% auf 80% an. Dargestellt sind die Wichtungsfaktoren w nach 336 Iterationen für die 82 Zielsequenzen, geordnet nach der Sequenzlänge.

	Trainingsset			
Testset	Set_{45}	Set_{135}	Set_{420}	Set_{1014}
Set_{45}	80%	89%	89%	89%
Set_{135}	68%	82%	79%	79%
Set_{420}	52%	72%	73%	74%
Set_{1014}	49%	64%	67%	67%

Tabelle 3.11: Erkennung der nativen Proteine. Die Energieparameter werden mit Hilfe der quasichemischen Methode, unter Verwendung von Gleichung 2.33 erzeugt. Es gilt das C_α Kontaktkriterium.

det, die beim Test nicht verwendet werden, so verbessert dies das Ergebnis beim Test. Die Übertragbarkeit der Energiefunktion ist relativ gut. Wird die Energiefunktion mit dem Set_{45} trainiert, so ist die Erkennung beim Set_{1014} immerhin noch 49%. Zur Erinnerung: Das Set_{1014} enthält $24.9 \cdot 10^6$ *Decoys* und 202 Zielsequenzen, das Set_{45} nur 29959 *Decoys* und 45 Zielsequenzen. Abb. 3.14 zeigt die Korrelation der Parameter generiert mit dem Set_{45} bzw. generiert mit dem Set_{1014}. Der Korrelationskoeffizient beträgt 0.81.

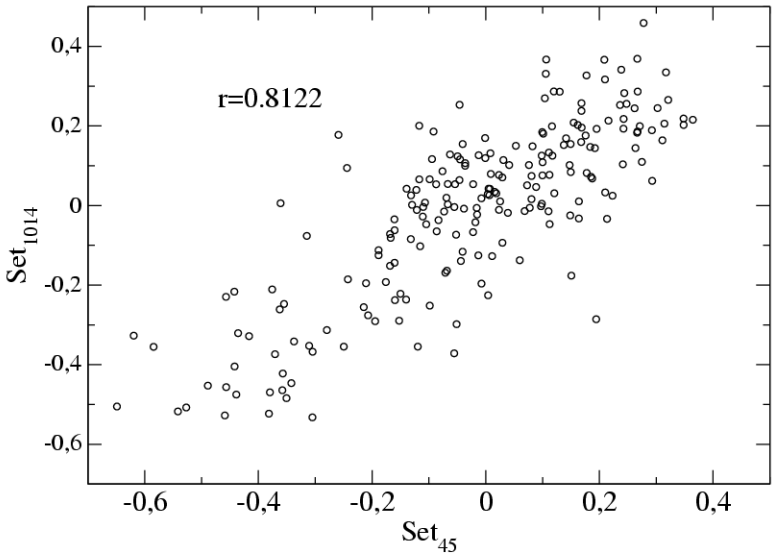

Abbildung 3.14: Korrelation der Kontaktenergieparameter erzeugt mit Set_{45}, bzw. mit Set_{1014} unter Verwendung der quasichemischen Methode nach Gleichung 2.35.

Bei Verwendung der quasichemischen Methode wird bei den *Decoys* hinsichtlich der Ähnlichkeit zur nativen Struktur nicht unterschieden. Eine Struktur kann nur entweder nativ oder nicht-nativ sein. Es scheint also angebracht zu sein *Decoys*, die eine gewisse Ähnlichkeit zur nativen Struktur haben, vom Lernen der Energiefunktion auszuschließen. Tabelle 3.12 zeigt die Erkennung der nativen Proteinstrukturen unter Verwendung verschiedener *Overlap* Grenzwerte q_{thr}. Die Parameter sind mittels der quasichemischen Methode bei gleicher Gewichtung aller Strukturen (Gleichung 2.33) erzeugt. Beim Lernen werden nur *Decoys* bis zu diesem *Overlap* verwendet, der Test der Energiefunktion hingegen erfolgt mit allen *Decoys*.
Aus Tabelle 3.12 ist ersichtlich, dass im Set_{135} mit einem Anteil von 5% nur sehr wenige Strukturen mit einem *Overlap* größer als 0.4 vorkommen. Werden alle *Decoys* mit einem *Overlap* $q > 0.4$ beim Lernen vernachlässigt, so verändert sich die Erkennung nicht und es werden auch die gleichen Proteine nicht erkannt. Wird ein Grenzwert

	Erkennung/ignorierte Strukturen		
Overlap Grenzwert q_{thr}	set_{135}	set_{420}	set_{1014}
0.2	54%/92%	39%/94%	29%/94%
0.3	74%/38%	69%/36%	63%/39%
0.4	82%/5%	75%/5%	67%/5%
0.5	82%/0.2%	73%/0.2%	67%/0.2%
0.6	82%/6	73%/525	67%/1230
0.7	82%/0	73%/10	67%/22
0.8	82%/0	73%/4	67%/6
0.9	82%/0	73%/3	67%/5
1.0	82%/0	73%/0	67%/0

Tabelle 3.12: Erkennung der nativen Proteine. Alle *Decoys* mit einem *Overlap* $q > q_{\text{thr}}$ werden beim Lernen der Energieparameter vernachlässigt. Das Lernen erfolgt mit Hilfe der quasichemischen Methode ohne Gewichtung (siehe Gleichung 2.33). Für den Test der Parameter werden alle *Decoys* verwendet.

von $q_{\text{thr}} = 0.2$ angelegt, so werden beim Lernen 92% aller Strukturen vernachlässigt. Trotzdem werden beim Test immerhin noch 54% aller nativen Strukturen als solche erkannt.

Erfolgt die Berechnung der Energieparameter mittels der quasichemischen Methode mit Gewichtung (siehe Gleichung 2.35), so steigt die Erkennung und auch die Übertragbarkeit stark an.

	Trainingsset			
Testset	Set_{45}	Set_{135}	Set_{420}	Set_{1014}
Set_{45}	96%	91%	89%	89%
Set_{135}	84%	83%	79%	80%
Set_{420}	77%	72%	74%	74%
Set_{1014}	69%	65%	69%	70%

Tabelle 3.13: Erkennung der nativen Proteine. Die Energieparameter wurden mit Hilfe der quasichemischen Methode unter Verwendung von Gleichung 2.35 bestimmt.

Wie aus Tabelle 3.13 ersichtlich ist, genügt es praktisch, die Energiefunktion nur mit Set_{45} zu trainieren. Die Erkennung für das Set_{1014} beträgt in diesem Fall 69%. Wird mit dem gesamten Set_{1014} gelernt, ist die Erkennung mit 70% nur um 1% besser. In

Abb. 3.15 werden die Energieparameter von Set_{45} und Set_{1014} verglichen. Die Korrelation ist mit einem Korrelationskoeffizienten von 0.9197 relativ hoch. Auffällig sind nur die Parameter für die Paare Cystein-Histidin und Histidin-Histidin. Hier haben die Parameter unterschiedliche Vorzeichen und unterscheiden sich für Cystein-Histidin im Betrag um 0.47 für Histidin-Histidin um 0.44[1]. Beim Set_{45} gibt es in den nativen Proteinen allerdings auch nur 7 Histidin-Histidin und 13 Histidin-Cystein Paare. Es ist zu vermuten, dass die schlechte Korrelation durch diese geringe Zahl bedingt ist. Zum Vergleich: Das Paar Glycin-Glycin tritt 239 mal auf. Abb. 3.16 zeigt die 210 Parameter für Set_{45}.
Werden beim Lernen *Decoys* mit hohem *Overlap q* zur nativen Struktur vernachlässigt, so wird die Erkennung interessanterweise nicht beeinträchtigt (siehe Tabelle 3.14). Bei Set_{135} und Set_{420} bewirkt ein *Overlap* Grenzwert von $q_{thr} = 0.2$ sogar eine Verbesserung gegenüber der Verwendung aller *Decoys*. Ein solcher Grenzwert bewirkt einen Ausschluss von 92% bzw. 94% aller *Decoys* beim Lernen.

	Erkennung/ignorierte Strukturen		
Overlap Grenzwert q_{thr}	set_{135}	set_{420}	set_{1014}
0.2	87%/92%	77%/94%	70%/94%
0.3	83%/38%	74%/36%	70%/39%
0.4	83%/5%	74%/5%	70%/5%
0.5	83%/0.2%	74%/0.2%	70%/0.2%
0.6	83%/6	74%/525	70%/1230
0.7	83%/0	74%/10	70%/22
0.8	83%/0	74%/4	70%/6
0.9	83%/0	74%/3	70%/5
1.0	83%/0	74%/0	70%/0

Tabelle 3.14: Erkennung der nativen Proteine. Alle *Decoys* mit einem *Overlap* $q > q_{thr}$ werden beim Lernen der Energieparameter vernachlässigt. Das Lernen erfolgt mit Hilfe der quasichemischen Methode mit Gewichtung (siehe Gleichung 2.35). Für den Test der Parameter werden alle *Decoys* verwendet.

all atom **Modell** Wird das *all atom* Modell angewandt, so steigt die Erkennung stark an (siehe Tabelle 3.15). Die Werte erreichen das gleiche Niveau wie bei der linearen Optimierung (siehe 2.10).

[1] Für beide Energiefunktionen gilt: $\sum_{i=1}^{210} u_i^2 = 10$.