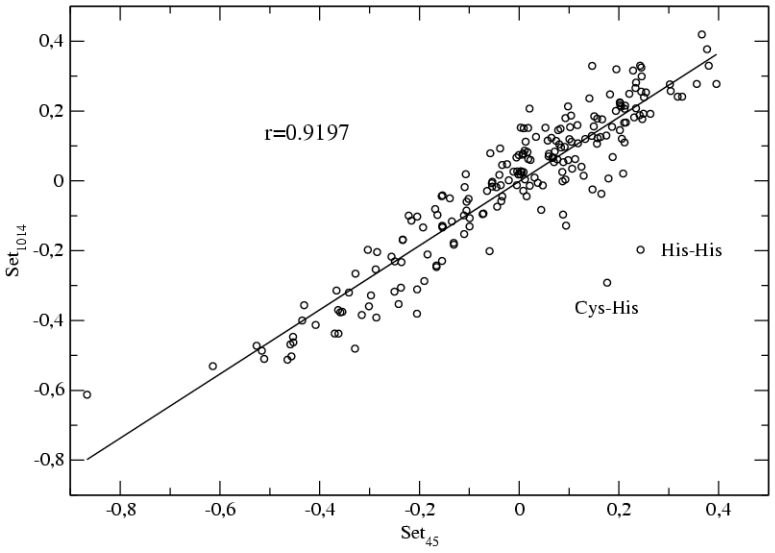

Abbildung 3.15: Die Kontaktenergieparameter berechnet mit Hilfe der quasichemischen Methode mit Gewichtung (siehe Gleichung 2.35). Das Training erfolgt einmal mit Hilfe von Set_{45} und einmal mit Hilfe von Set_{1014}. Die Korrelation zwischen den beiden Energiefunktion ist mit r=0.9197 relativ hoch, wobei die Parameter für Histidin-Histidin und Cystein-Histidin relativ stark voneinander abweichen.

	Set_{135}	Set_{420}	Set_{1014}
ohne Gewichtung	98%	94%	94%
mit Gewichtung	96%	96%	95%

Tabelle 3.15: Erkennung von nativen Proteinen unter Verwendung des *all atom* Modells. Die Energieparameter werden mit Hilfe der quasichemischen Methode ohne Gewichtung (Gleichung 2.33) bzw. mit Gewichtung (Gleichung 2.35) erzeugt.

3.3.2 Der Kontaktabstand

Wichtig für einen erfolgreichen Einsatz einer Kontaktenergiefunktion ist die richtige Wahl der Parameter. Die Energieparameter der Kontakte werden über verschiedene Methoden berechnet. Die verbleibenden Parameter, wie z.B. Kontakt- und Sequenzabstand (siehe 2.2.1), müssen jedoch auch in sinnvoller Weise verwendet werden. Abb. 3.17 zeigt die Verteilungen der C_α-C_α Abstände für ausgewählte Aminosäurepaare. Um eine gute Statistik zu erhalten sind hierbei alle Proteine aus Set_{1014} verwendet. Berücksichtigt sind nur Abstände zwischen Residuen mit einem Sequenzabstand von mindestens $dis_{seq} = 3$. Wie man sieht, unterscheiden sich die Verteilungen nur geringfügig. Die Verteilungen der Paare Glycin-Glycin und Tryptophan-Tryptophan sind sehr

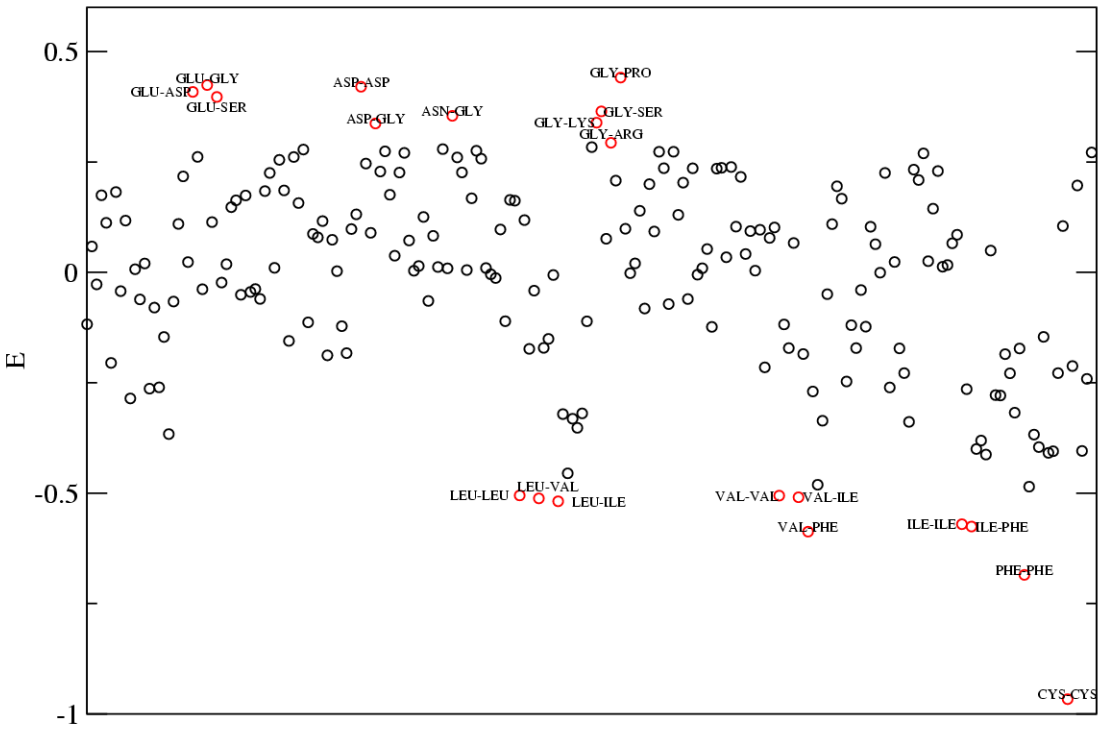

Abbildung 3.16: Die 210 Kontaktenergieparameter bestimmt nach der quasichemischen Methode mit Gewichtung (siehe Gleichung 2.35). Das Training erfolgt mit Hilfe von Set_{45}. Für Energieparameter mit besonders hohen Beträgen ist das zugehörige Aminosäurepaar angegeben.

ähnlich, obwohl sich die Aminosäuren in ihrer Größe stark unterscheiden.

Somit ist auch für die Verwendung von unterschiedlichen Kontaktabständen für unterschiedliche Aminosäurepaare nur eine geringfügige Verbesserung bei der Erkennung zu erwarten.

Tabelle 3.16 zeigt die Erkennungen für verschiedene Kontaktabstände unter Verwendung der verschiedenen Methoden. Die optimalen Kontaktabstände von NMR und Kristallstrukturen unterscheiden sich bei der linearen Optimierung deutlich voneinander. So erweist sich ein Kontaktabstand von 11Å für Kristallstrukturen als ein sinnvoller Wert. Bei den NMR Strukturen ergibt ein Abstand von 8Å bei der linearen Optimierung die beste Erkennung.

3.3.3 Verwendung von verschiedenen Sequenzabstandsbereichen

Wie in 2.2.2 angedeutet, kann die Kontaktenergiefunktion erweitert werden, indem verschiedene Sequenzabstandsbereiche unterschieden werden. In α Helices sind vor

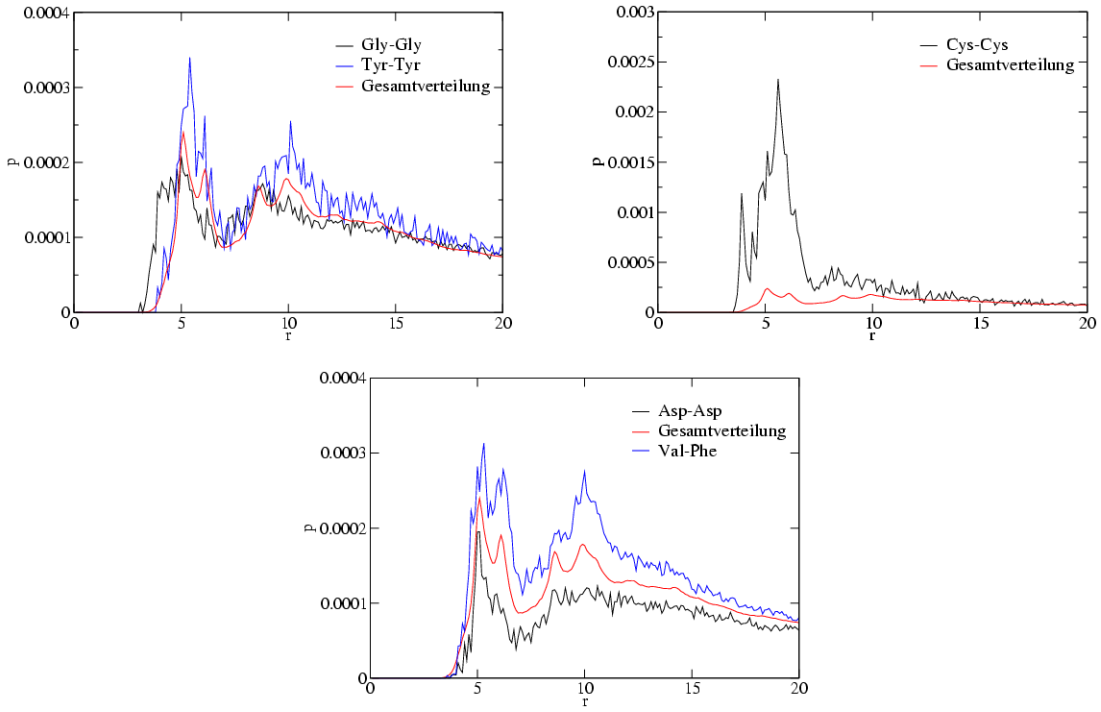

Abbildung 3.17: Abstandsverteilungen für ausgewählte Aminosäurepaare im Vergleich zur Abstandsverteilung aller Aminosäurepaare. Um eine gute Statistik zu erhalten sind alle Proteinstrukturen aus Set_{1014} verwendet. Aufgetragen ist die normalisierte Häufigkeit der C_α-C_α Abstände gegen r: $f(r) = \frac{N(r,\Delta r)}{r^2 \Delta r}$

Kontaktabstand	6Å	7Å	8Å	9Å	10Å	11Å	12Å	13Å	14Å	15Å
LO Set_{135}	26%	50%	62%	60%	68%	70%	65%	60%	59%	50%
QCM Set_{135}	41%	63%	68%	72%	78%	82%	83%	80%	72%	70%
QCMw Set_{135}	33%	55%	70%	71%	76%	83%	80%	80%	74%	71%
QCMw Set_{420}	29%	47%	66%	65%	69%	74%	73%	70%	67%	63%
LO Set_{NMR}	33%	53%	54%	52%	52%	48%	48%	41%	30%	27%
QCMw Set_{NMR}	38%	53%	64%	64%	65%	67%	64%	60%	59%	54%

Tabelle 3.16: Erkennung von nativen Proteinen in Abhängigkeit vom Kontaktabstand r_c (siehe 2.2.1) unter Verwendung von verschiedenen Methoden zur Bestimmung der Kontaktenergieparameter. Die Abkürzungen finden sich in 6.3.

allem (i,i+3) und (i,i+4) Wechselwirkungen von Relevanz. Es erscheint daher sinnvoll, diesen Bereich gesondert zu betrachten. Es wird also unterteilt in Kontakte zwischen Residuen im Abstand (i,i+3) oder (i,i+4) und allen höheren Sequenzabständen. Es zeigt sich, dass sich für das Set_{135} die Erkennung hierbei nicht verändert. Sie ist, genau wie für das Modell mit nur einem Abstandsbereich, 70%. Die Parameter für die beiden Bereiche sind nahezu unkorreliert (siehe Abb. 3.18). Für Set_{420} erhöht sich die Erkennung von 52% auf 59%.

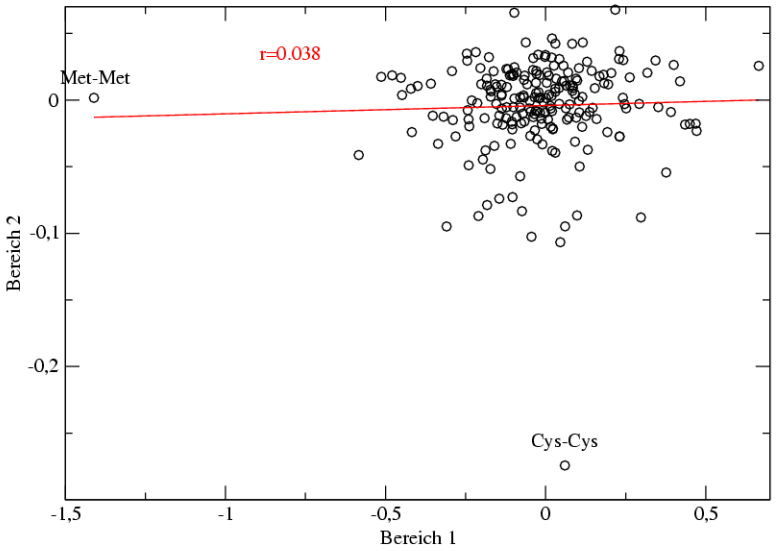

Abbildung 3.18: Die Energieparameter für Sequenzabstände von drei oder vier (Bereich 1) sowie für Sequenzabstände größer als vier (Bereich 2). Die Parameter sind praktisch unkorreliert. Die Energiefunktion wurde mit Hilfe der linearen Optimierung (siehe 2.10) unter Verwendung von Set_{135} generiert.

Tabelle 3.17 zeigt die Erkennung für verschiedene Abstandsbereiche. Von den betrachteten Bereichen ist die Erkennung mit 85% am größten für die Unterteilung in die Bereiche drei bis sieben sowie alle Abstände größer als sieben.

3.3.4 Ein zusätzlicher Parameter für Aminosäuren an der Proteinoberfläche

Werden mit Hilfe von Formel 2.13 Wechselwirkungen der Proteinoberfläche mit der Umgebung in die Energiefunktion mit einbezogen, so ist die Wahl des Parameters A (minimale Zahl der Kontakte für individuelle Residuen) von entscheidender Bedeutung. Da die Zahl der Kontakte für einzelne Aminosäuren stark streut, ist es schwierig einen sinnvollen Wert für A direkt aus den Daten abzuleiten. Die Erkennungen für

	Erkennung	
Abstandsbereiche	Set_{135}	Set_{420}
3-4, >4	70%	59%
3-5, >5	72%	59%
3-6, >6	79%	61%
3-7, >7	85%	65%
3-8, >8	82%	60%
3-9, >9	78%	55%

Tabelle 3.17: Wird 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 der nativen Proteine verbessern. Die verwendete Energiefunktion wurde mit Hilfe der linearen Optimierung (siehe 2.10) unter Verwendung von Set_{135} generiert.

Set_{135} unter Verwendung von unterschiedlichen Werten für A sind in Abb. 3.19 dargestellt.

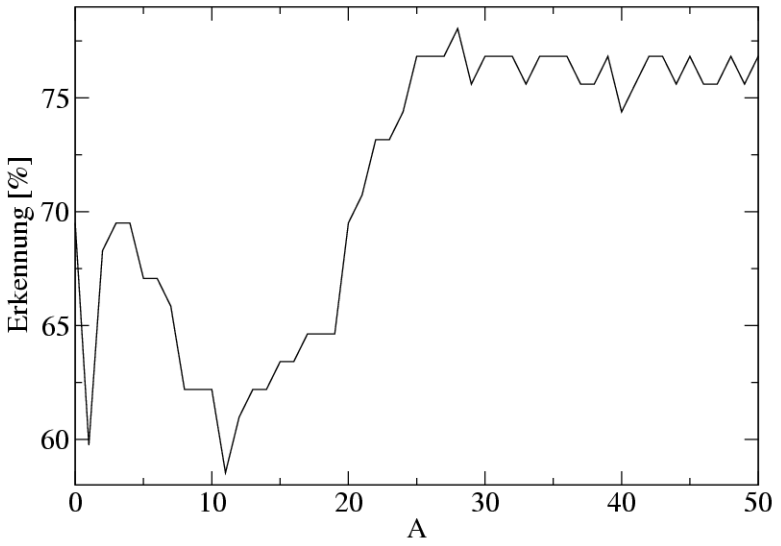

Abbildung 3.19: Wechselwirkungen der Proteinoberfläche mit Lösungsmittelmolekülen können durch Hinzunahme von Kontaktenergieparametern für Kontakte zwischen Residuen und dem Lösungsmittel berücksichtigt werden. Hierfür wird angenommen, dass eine Aminosäure mit Lösungsmittelmolekülen wechselwirkt, wenn sie weniger als eine vorgegebene Anzahl A an Nachbarn besitzt. Dargestellt ist die Erkennung für Set_{135} in Abhängigkeit von diesem Oberflächenparameter A (siehe Gleichung 2.13).

Ab $A = 25$ schwankt die Erkennung um einen Wert von 77%. Für die Kontaktenergiefunktion ohne Oberflächenenergieterme ist die Erkennung 70%. Es liegt also eine

beträchtliche Steigerung vor, wenn man berücksichtigt, dass die Zahl der Kontaktparameter nur um 20 von 210 auf 230 erhöht wird. Abb. 3.20 zeigt die 20 neuen Energieparameter für $A = 27$.

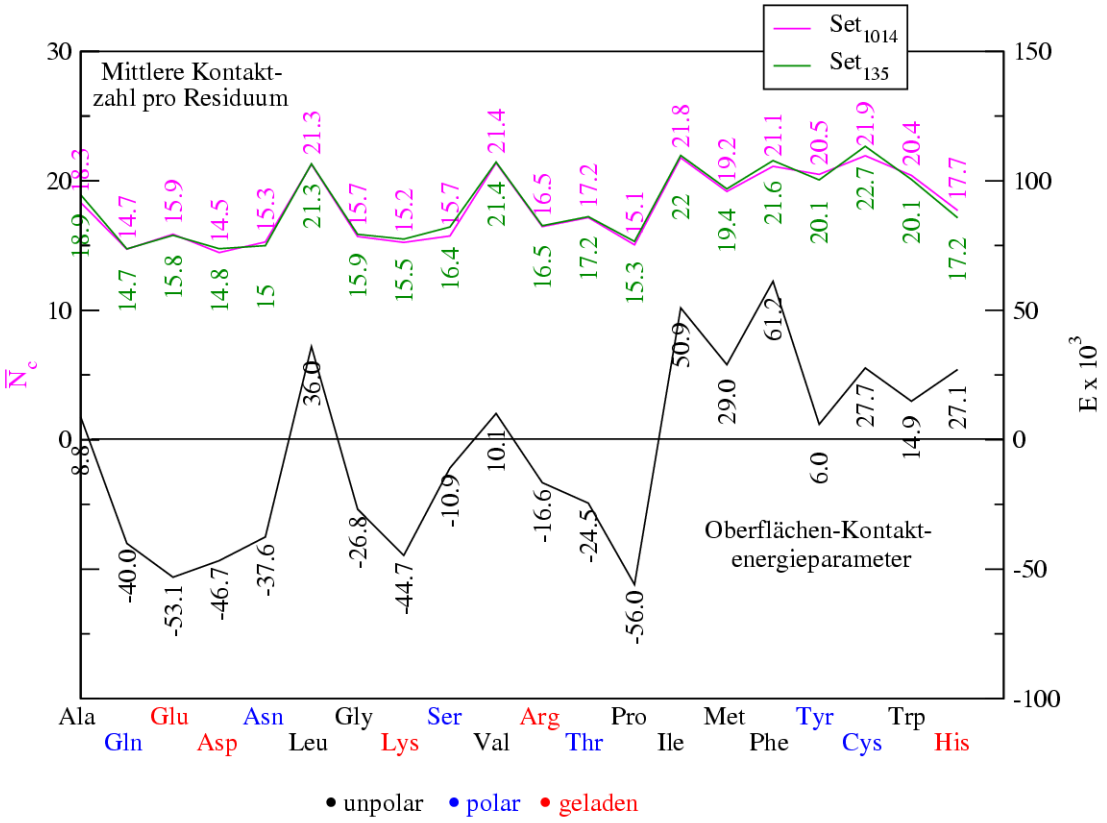

Abbildung 3.20: Die zwanzig Energieparameter für Kontakte mit der Umgebung (schwarze Kurve) (siehe Gleichung 2.13) sowie die mittleren Kontaktzahlen für die verschiedenen Aminosäuren für die 82 Zielsequenzen aus Set_{135} (grüne Kurve), bzw. die 202 Zielsequenzen aus Set_{1014} (Linie in Magenta).

Die Parameter korrelieren gut mit den mittleren Kontaktzahlen der verschiedenen Aminosäuren. Mit Ausnahme von Histidin haben sämtliche Parameter für Aminosäuren mit $\overline{N}_c < 18$ ein negatives Vorzeichen. Sämtliche Parameter für Aminosäuren mit $\overline{N}_c > 18$ haben ein positives Vorzeichen.
Histidin tritt häufig als axialer Ligand von Häm-Gruppen auf. Bei der Darstellung einer Proteinstruktur als Kontaktmatrix werden Häm-Gruppen nicht berücksichtigt (siehe 2.2.5). Bereiche die in einem Protein von Häm-Gruppen besetzt sind, werden also von der Kontaktmatrix als Lücke wiedergegeben. Die Oberflächen Kontaktenergieparameter in Abb. 3.20 sind mit Set_{135} bestimmt. Die 82 Zielsequenzen in diesem Set enthalten keine Häm-Gruppen (siehe 2.7). Die mittleren Kontaktzahlen der 20 Aminosäuren unterscheiden sich nur geringfügig für Set_{135} und Set_{1014}, Histidin hat in Set_{135}

sogar eine geringfügig kleinere Zahl an Kontakten als in Set_{1014}. Häm-Gruppen sind also nicht der Grund für die Diskrepanz zwischen mittlerer Kontaktzahl und Oberflächen Kontaktenergieparameter von Histidin.

3.3.5 Erkennen von Strukturen mit nativen Eigenschaften

Neben der Erkennung von nativen Strukturen ergibt sich die Frage, ob Strukturen die zwar einen *Overlap* kleiner eins aufweisen jedoch native Eigenschaften im Sinne eines hohen *Overlaps* nahe eins besitzen, ebenfalls erkannt werden.
Wie in Kapitel 2.12 angedeutet, kann man die mittels *Threading* für eine Sequenz erzeugten Strukturen verwenden, um eine Strukturvorhersage zu machen. Tabelle 3.18 zeigt für die verschiedenen Sequenzen jeweils den *Overlap* der auf diese Weise erhaltenen energieärmsten nicht-nativen Struktur ($q_{\min}$). Desweiteren ist angegeben der höchste *Overlap*, der für diese Sequenz unter allen *Decoys* zu finden ist ($q_{\max}$).
Im Set_{1014} haben 14 Zielsequenzen einen *Decoy* mit $q_{\max} \geq 0.7$. Tabelle 3.18 gibt für diese Sequenzen den Wert $q_{\max}$ an sowie für verschiedene Methoden zur Generierung der Energieparameter den *Overlap* des *Decoys* mit der niedrigsten Energie $q_{\min}$. Zur Anwendung kommt hier die lineare Optimierung nach 2.24 unter Verwendung eines Polynoms sowie die quasichemische Methode mit Gewichtung (siehe Gleichung 2.35). Zum Lernen der Energieparameter werden die vier verschiedenen Kristallstruktur-Proteinsets aus 2.7 verwendet.
Verwendet man bei der quasichemischen Methode mit Gewichtung (siehe Gleichung 2.35) für das Lernen nicht alle *Decoys*, sondern nur die mit einem *Overlap* q kleiner gleich einem *Overlap*-Grenzwert $q_{\text{thr}} = 0.2$, so ist die Erkennung der nativen Strukturen für Set_{1014} genauso erfolgreich, als wenn alle *Decoys* für das Lernen verwendet werden. Für Set_{135} und Set_{420} ist die Erkennung sogar erfolgreicher als bei Verwendung aller *Decoys*. Hier wird getestet, wie gut diese Energiefunktionen bei der Erkennung von *Decoys* mit hohen q-Werten funktionieren.
Bei sieben Sequenzen liegt bei keiner Methode eine Vorhersage mit einem *Overlap* größer 0.5 vor. Für die Sequenz 2cpl wird der ähnlichste *Decoy* ($q_{\max} = 0.96$) unter Verwendung der quasichemischen Methode erkannt wenn das Training mit Set_{45} erfolgt, bzw. wenn Set_{45}, Set_{135}, oder Set_{420} verwendet wird und das Training nur mit *Decoys* mit einem *Overlap* $q \leq 0.2$ erfolgt. Auffällig hierbei ist, dass bei Verwendung aller *Decoys* beim Lernen der Energiefunktion nur bei einem Training mit dem Set_{45} der ähnlichste *Decoy* erkannt wird. Werden nur *Decoys* mit einem *Overlap* $q \leq 0.2$ aus Set_{45}, Set_{135} oder Set_{420} für das Training verwendet, so wird für die Sequenz 2cpl der *Decoy* mit dem maximalem *Overlap* $q_{\max}$ erfolgreich erkannt. Mit Set_{1014} erfolgt

weder nach Training mit allen *Decoys* noch bei Verwendung von $q_{\mathrm{thr}} = 0.2$ eine Erkennung. Ein Training mit weniger Strukturen führt hier also zu einer besseren Erkennung. Beim Training über die quasichemische Methode sind für die Sequenz 2cpl die kleineren Sets erfolgreicher. Beim Training über die lineare Optimierung ist der Fall umgekehrt: das Training mit dem kleinsten Set führt zu falscher Erkennung, während mit allen anderen Sets die Erkennung erfolgreich ist.

Die beste Vorhersage für Sequenz 1flp hat einen *Overlap* q_{min} von 0.51, für den ähnlichsten *Decoy* gilt $q_{\mathrm{max}} = 0.71$. Die quasichemische Methode liefert für diese Sequenz Vorhersagen mit einem *Overlap* q_{min} zwischen 0.31 und 0.51, während die lineare Optimierung *Overlaps* zwischen 0.29 und 0.44 liefert. Für die Sequenz 1eca wird der ähnlichste *Decoy* mit $q_{\mathrm{max}} = 0.70$ nicht erkannt, sondern von der quasichemischen Methode ein *Decoy* mit $q_{\mathrm{min}} = 0.62$ vorgeschlagen. Die lineare Optimierung erreicht für diese Sequenz nur $q_{\mathrm{min}} = 0.38$. Wieder sind es bei der quasichemischen Methode eher die kleinen Trainingssets, die erfolgreich sind.
Die Sequenz 1tmy stellt für alle drei Typen von Energiefunktion kein Problem dar. Mit Ausnahme von $\mathrm{QCM}_{0.2/45}$ wird der ähnlichste *Decoy* ($q = 0.77$) von allen Energiefunktionen erkannt. Bei der Sequenz 2pvb sind die optimierten Energiefunktionen nicht erfolgreich, bei der quasichemischen Methode hingegen ist nur die Energiefunktion $\mathrm{QCMw}_{0.2/45}$ nicht in der Lage den ähnlichsten *Decoy* zu erkennen. Für die Sequenz 1r69 liefert die mittels linearer Optimierung und Set_{45} generierte Energiefunktion einen *Overlap* von immerhin $q_{\mathrm{min}} = 0.59$, mit Set_{135} ergibt sich $q_{\mathrm{min}} = 0.52$, alle übrigen Energiefunktionen liefern q_{min}-Werte von unter 0.5. Die Sequenz 2erl wird von den mit Hilfe der quasichemischen Methode abgeleiteten Energiefunktionen im Mittel etwas besser vorhergesagt als von den optimierten Energiefunktionen.

Korrelation zwischen q_{min} und dem *Z-Score* Die *Overlaps* der energieärmsten *Decoys* q_{min} fallen nicht immer mit dem höchsten *Overlap* aller *Decoys* q_{max} zusammen. Es wäre also hilfreich ein Kriterium anwenden zu können, welches erlaubt, die Qualität der Vorhersage abzuschätzen.
Hierbei bietet sich z.B. der *Z-Score* an, welcher die Stabilität einer Struktur relativ zum Mittelwert aller betrachteten Strukturen angibt (siehe Gleichung 2.37). Tabelle 3.19 zeigt die Korrelationskoeffizienten zwischen q_{min} und dem *Z-Score* für die verschiedenen Energiefunktionen. Die stärksten Korrelationen ergeben sich für die lineare Optimierung (LO). Abb. 3.21 zeigt für LO Set_{135}, LO Set_{420} und LO Set_{1014} die Werte q_{min} in Abhängigkeit vom *Z-Score* für die 14 verschiedenen Sequenzen. Diese drei

Energiefunktionen erkennen jeweils für die gleichen zwei der 14 Proteinsequenzen korrekt den ähnlichsten *Decoy*. Für LO Set_{135} sind die *Z-Scores* zu drei Sequenzen deutlich abgesenkt. Die beiden erkannten Sequenzen sind hierin enthalten. Wird als Energiefunktion LO Set_{420} verwendet, so weisen die beiden erkannten Sequenzen die niedrigsten Werte für den *Z-Score* auf. Unter Verwendung der Energiefunktion LO Set_{1014} sind die *Z-Scores* zu den zwei erkannten sowie zu zwei weiteren Sequenzen deutlich erniedrigt.
Es zeigt sich also, dass zwar nur eine geringe Zahl an ähnlichsten *Decoys* erkannt wird, jedoch eine starke Korrelation zwischen *Z-Score* und *Overlap* des jeweils energieärmsten *Decoys* besteht. Eine solche Eigenschaft ist sehr nützlich: Es liegt zwar nur in wenigen Fällen eine sinnvolle Vorhersage vor, jedoch kann abgeschätzt werden, welche Vorhersagen sinnvoll sind.

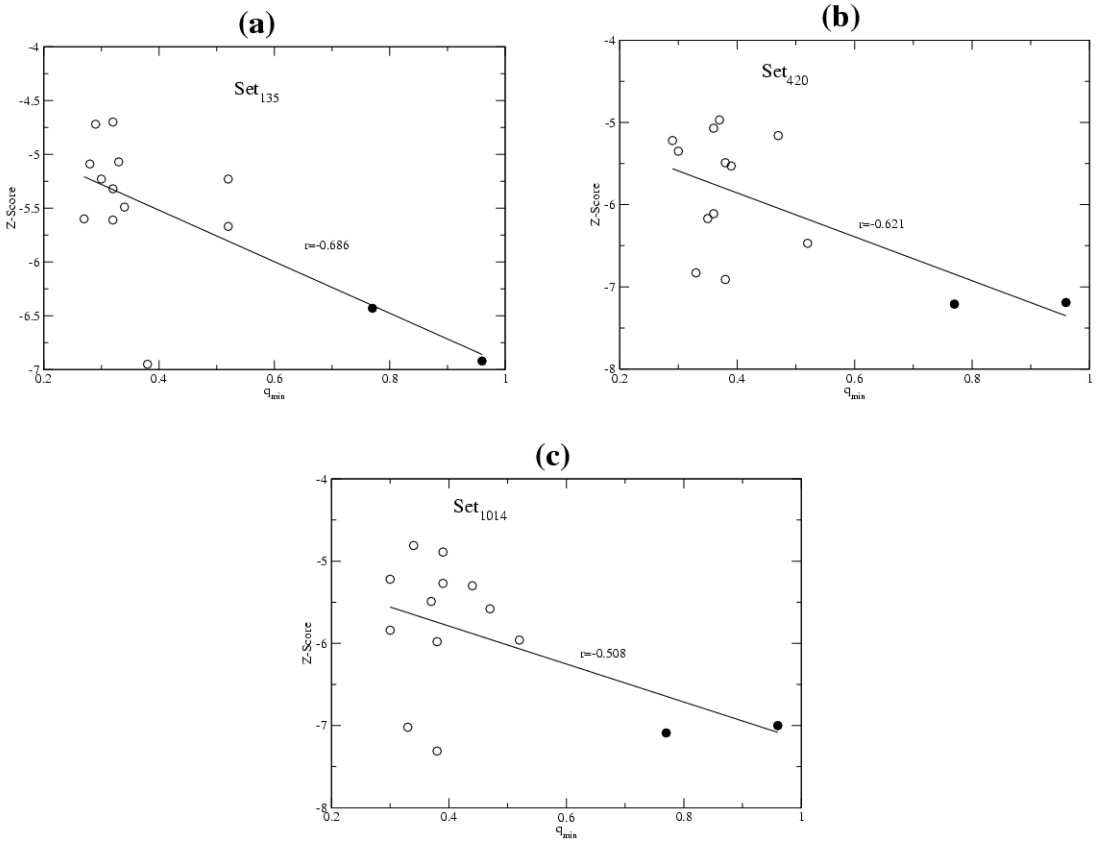

Abbildung 3.21: Korrelation zwischen dem *Overlap* des *Decoys* geringster Energie q_{min} und dem Z-Score (siehe Gleichung 2.37) für die 14 Sequenzen aus Set_{1014} mit größtem q_{max}. Dargestellt sind die Korrelationen bei Verwendung der linearen Optimierung (LO). Als Lernset der Energiefunktion diente Set_{135} (a), Set_{420} (b), bzw. Set_{1014} (c). •: Der ähnlichste *Decoy* wird erkannt. ○: Der ähnlichste *Decoy* wird nicht erkannt.

		1aew	2cpl	1flp	1eca	1lid
q_{max}		0.96	0.96	0.71	0.70	0.78
q_{min}/*Z-Score*	QCMw Set_{45}	0.32/-4.98	0.96/-3.43	0.47/-4.04	0.62/-4.72	0.36/-5.49
	QCMw Set_{135}	0.32/-4.44	0.21/-3.37	0.47/-3.66	0.62/-4.32	0.30/-4.52
	QCMw Set_{420}	0.32/-4.86	0.26/-3.46	0.51/-3.71	0.52/-4.46	0.36/-4.58
	QCMw Set_{1014}	0.32/-4.92	0.26/-3.42	0.51/-3.77	0.52/-4.39	0.36/-4.60
	$QCMw_{0.2/45}$	0.21/-3.26	0.96/-3.64	0.31/-3.38	0.28/-3.61	0.28/-3.31
	$QCMw_{0.2/135}$	0.32/-4.81	0.96/-3.63	0.47/-4.08	0.62/-4.72	0.36/-5.44
	$QCMw_{0.2/420}$	0.32/-5.06	0.96/-3.52	0.51/-3.93	0.52/-4.52	0.36/-5.11
	$QCMw_{0.2/1014}$	0.32/-4.98	0.26/-3.40	0.51/-3.78	0.52/-4.36	0.36/-4.77
	LO Set_{45}	0.32/-4.82	0.27/-5.27	0.30/-4.95	0.38/-5.20	0.36/-5.19
	LO Set_{135}	0.32/-5.32	0.96/-6.92	0.29/-4.72	0.28/-5.09	0.30/-5.23
	LO Set_{420}	0.29/-5.22	0.96/-7.19	0.37/-4.97	0.36/-5.07	0.36/-6.11
	LO Set_{1014}	0.34/-4.81	0.96/-7.00	0.44/-5.30	0.37/-5.49	0.30/-5.84

		1tmy	2pvb	1erv	1aac	1opd
q_{max}		0.77	0.93	0.73	0.96	0.89
q_{min}/*Z-Score*	QCMw Set_{45}	0.77/-5.91	0.93/-5.11	0.38/-5.75	0.30/-5.30	0.23/-4.76
	QCMw Set_{135}	0.77/-5.81	0.93/-4.39	0.38/-4.92	0.30/-4.45	0.31/-4.03
	QCMw Set_{420}	0.77/-5.92	0.93/-4.26	0.38/-4.82	0.30/-4.70	0.31/-4.07
	QCMw Set_{1014}	0.77/-5.99	0.93/-4.26	0.38/-4.80	0.30/-4.68	0.31/-4.03
	$QCMw_{0.2/45}$	0.30/-3.29	0.35/-3.55	0.34/-3.48	0.29/-3.23	0.24/-3.69
	$QCMw_{0.2/135}$	0.77/-6.20	0.93/-4.99	0.38/-5.35	0.30/-5.00	0.31/-4.46
	$QCMw_{0.2/420}$	0.77/-6.27	0.93/-4.53	0.38/-5.10	0.30/-5.00	0.31/-4.26
	$QCMw_{0.2/1014}$	0.77/-6.05	0.93/-4.29	0.38/-4.84	0.30/-4.72	0.31/-4.08
	LO Set_{45}	0.77/-5.28	0.37/-5.32	0.38/-5.95	0.33/-4.58	0.28/-4.96
	LO Set_{135}	0.77/-6.43	0.32/-5.61	0.38/-6.95	0.33/-5.07	0.27/-5.60
	LO Set_{420}	0.77/-7.21	0.33/-6.83	0.38/-6.91	0.30/-5.35	0.35/-6.17
	LO Set_{1014}	0.77/-7.09	0.33/-7.02	0.38/-7.31	0.30/-5.22	0.38/-5.98

		1r69	1vie	2fdn	2erl
q_{max}		0.74	0.99	0.80	0.71
q_{min}/*Z-Score*	QCMw Set_{45}	0.43/-4.14	0.33/-4.23	0.43/-5.21	0.55/-5.90
	QCMw Set_{135}	0.30/-3.39	0.35/-3.96	0.39/-5.23	0.55/-5.70
	QCMw Set_{420}	0.30/-3.40	0.29/-4.30	0.40/-4.97	0.55/-5.54
	QCMw Set_{1014}	0.26/-3.42	0.29/-4.34	0.40/-4.96	0.53/-5.54
	$QCMw_{0.2/45}$	0.41/-4.06	0.37/-3.80	0.42/-3.85	0.45/-4.90
	$QCMw_{0.2/135}$	0.43/-4.14	0.35/-4.30	0.39/-5.28	0.55/-5.68
	$QCMw_{0.2/420}$	0.26/-3.39	0.29/-4.49	0.40/-5.04	0.55/-5.67
	$QCMw_{0.2/1014}$	0.26/-3.41	0.29/-4.35	0.40/-4.97	0.55/-5.57
	LO Set_{45}	0.59/-4.69	0.36/-4.79	0.43/-5.04	0.44/-5.68
	LO Set_{135}	0.52/-5.23	0.32/-4.70	0.34/-5.49	0.52/-5.67
	LO Set_{420}	0.47/-5.16	0.38/-5.49	0.39/-5.53	0.52/-6.47
	LO Set_{1014}	0.47/-5.58	0.39/-4.89	0.39/-5.27	0.52/-5.96

Tabelle 3.18: q_{max}: Größte *Overlaps* aller *Decoys*, die mit dem Set_{1014} für die verschiedenen Sequenzen mittels *Threading* generiert werden. q_{min}: die *Overlaps* der jeweils energieärmsten *Decoys* für die 14 Sequenzen aus Set_{1014} mit größtem q_{max} unter Verwendung dreier Typen von Energiefunktionen. Desweiteren ist für die energieärmsten *Decoys* jeweils der *Z-Score* (siehe Gleichung 2.37) angegeben.

	r		r
QCMw Set_{45}	0.157	$QCMw_{0.2/420}$	-0.010
QCMw Set_{135}	-0.454	$QCMw_{0.2/1014}$	-0.400
QCMw Set_{420}	-0.386	LO Set_{45}	-0.095
QCMw Set_{1014}	-0.400	LO Set_{135}	-0.686
$QCMw_{0.2/45}$	-0.284	LO Set_{420}	-0.621
$QCMw_{0.2/135}$	0.046	LO Set_{1014}	-0.508

Tabelle 3.19: Korrelationskoeffizienten r zwischen dem *Overlap* des *Decoys* geringster Energie $q_{\min}$ und dem *Z-Score* (siehe Gleichung 2.37) für die 14 Sequenzen aus Set_{1014} mit größtem $q_{\max}$.

Sequenzähnlichkeiten Eine wichtige Frage ist, inwieweit die Sequenz einer Struktur, die für die Erzeugung eines *Decoys* verwendet wird, mit der Zielsequenz übereinstimmt.
Wird von allen verfügbaren *Decoys* der *Decoy* mit der höchsten Sequenzübereinstimmung gefunden, so wäre die Erkennung auch einfach alleine über diese Information möglich gewesen. Die Sequenzübereinstimmungen für die ähnlichsten *Decoys* für die 14 Sequenzen aus Tabelle 3.18 finden sich in Tabelle 3.20.

	1aew	2cpl	1flp	1eca	1lid	1tmy	2pvb
$q_{\max}$	0.96	0.96	0.71	0.70	0.78	0.77	0.93
Sequenzähnlichkeit	0.54	1.00	0.13	0.16	0.26	0.27	0.45
maximale Sequenzähnlichkeit	0.54	1.00	0.17	0.19	0.26	0.27	0.45
	1erv	1aac	1opd	1r69	1vie	2fdn	2erl
$q_{\max}$	0.73	0.96	0.89	0.74	0.99	0.80	0.71
Sequenzähnlichkeit	0.12	0.92	0.34	0.21	1.00	0.31	0.03
maximale Sequenzähnlichkeit	0.20	0.92	0.34	0.22	1.00	0.31	0.30

Tabelle 3.20: Ähnlichkeiten zwischen Zielsequenzen und Sequenzen die für die Erzeugung des *Decoys* mit größtem *Overlap* $q_{\max}$ verwendet werden. Desweiteren ist die maximal vorhandene Ähnlichkeit unter allen verwendeten *Decoys* angegeben.

Bei dem Protein mit dem PDB Code 2cpl handelt es sich um humanes Cyclophilin A. Beim ähnlichsten *Decoy* handelt es sich um das selbe Protein, welches jedoch als Komplex mit einem kurzen Peptid vorliegt und sich daher in der Struktur geringfügig von freiem Cyclophilin A unterscheidet.

Zum Protein 1flp gibt es eine Sequenz mit einer Ähnlichkeit von 0.17. Sequenzen mit höherer Ähnlichkeit liegen nicht vor. Der *Decoy* mit größtem *Overlap* ($q_{\max} = 0.71$) weist eine Sequenzähnlichkeit von 0.13 auf.

all atom **Modell** Tabelle 3.21 zeigt die Ähnlichkeiten der energieärmsten *Decoys* zu den entsprechenden nativen Kontaktmatrizen unter Verwendung des *all atom* Kriteriums.

	1aew	2cpl	1flp	1eca	1lid	1tmy	2pvb
$q_{\max}$	0.91	0.94	0.65	0.64	0.58	0.66	0.90
$q_{\min}$	0.91	0.94	0.61	0.64	0.08	0.66	0.90
	1erv	1aac	1opd	1r69	1vie	2fdn	2erl
$q_{\max}$	0.58	0.92	0.87	0.62	0.99	0.63	0.71
$q_{\min}$	0.24	0.92	0.87	0.36	0.15	0.20	0.37

Tabelle 3.21: Größter *Overlap* $q_{\max}$ aller *Decoys* sowie *Overlap* $q_{\min}$ des energieärmsten *Decoys* für die 14 Sequenzen aus Tabelle 3.18 unter Verwendung des *all atom* Kontaktkriteriums. Die Energiefunktion wurde mit Hilfe der linearen Optimierung (siehe 2.10) unter Verwendung von Set_{1014} generiert.

3.4 Monte Carlo Simulationen

Bei Verwendung einer sinnvollen Energiefunktion sollte sich eine gegebene native Struktur während einer Monte Carlo Simulation bei niedrigen Temperaturen nicht zu stark ändern.
Abb. 3.22 zeigt die native Struktur von Crambin sowie die Struktur nach einer Simulation mit 10^6 Schritten bei einer Temperatur von 0K. Die Kontaktenergieparameter sind mit der linearen Optimierung nach Gleichung 2.24 mit einem Polynom in $(1-q)$ vom Grade $\Delta = 5$ (siehe 2.25) unter Verwendung des Sets_{420} erzeugt. Crambin ist in Set_{420} nicht enthalten. Die Akzeptanzrate der Monte Carlo Simulation beträgt 0.04, die Trajektorie enthält also rund 40000 Strukturen. Der *Overlap* nach der MC Simulation beträgt $q = 0.78$, die C_α cRMSD beträgt 4.44Å. Struktur und Kontaktmatrix vor und nach der MC Simulation sind in Abb. 3.22 dargestellt. Wie man sieht sind die beiden α Helices weitestgehend erhalten. Das β Faltblatt ist zwar in der Strukturdarstellung nicht mehr als solches zu erkennen, die Kontakte zwischen den entsprechenden Residuen sind jedoch noch vorhanden (rote Kontakte in Abb. 3.22c). Abb. 3.23 zeigt den *Overlap* der letzten Struktur der MC Trajektorie mit der nativen Struktur von Crambin in Abhängigkeit von der Temperatur. Werden die Kontaktenergieparameter mit Hilfe der quasichemischen Methode ohne Gewichtung (siehe Gleichung 2.33) unter Verwendung von Set_{420} erzeugt, so ergibt sich bei ansonsten gleichen Simulationsbedingungen eine Akzeptanzrate von 0.03. Der *Overlap* der letzten Struktur der Monte Carlo Trajektorie zur nativen Struktur von Crambin beträgt 0.78 die C_α cRMSD 5.69Å.

3.4.1 Faltungssimulationen

Als elementarer Test für eine verwendete Energiefunktion werden die Kontaktenergieparameter nur für eine Sequenz trainiert. Die Monte Carlo Simulation erfolgt dann mit dieser Sequenz in einer gestreckten Konformation als Startpunkt.
Lässt sich auf diese Weise das Protein nicht zu einer nativ-ähnlichen Struktur falten, so ist nicht zu erwarten, dass eine echte Strukturvorhersage funktioniert.
Abb. 3.24 zeigt die Struktur von Crambin nach einer solchen Simulation. Die C_α cRMSD beträgt 4.93Å. Die Energiefunktion ist mittels der linearen Optimierung nach Gleichung 2.24 unter Verwendung eines Polynoms in $(1-q)$ (siehe 2.25) trainiert. Als Trainingsset dient Set_{420} zusammen mit der nativen Struktur von Crambin unter Verwendung von Crambin als einziger Zielsequenz. Die Faltungssimulation erfolgt bei einer Starttemperatur von 2000K. In einer *simulated annealing* Prozedur wird die Temperatur alle 10^5 Schritte mit dem Faktor 0.99 herunterskaliert. Insgesamt enthält die Si-

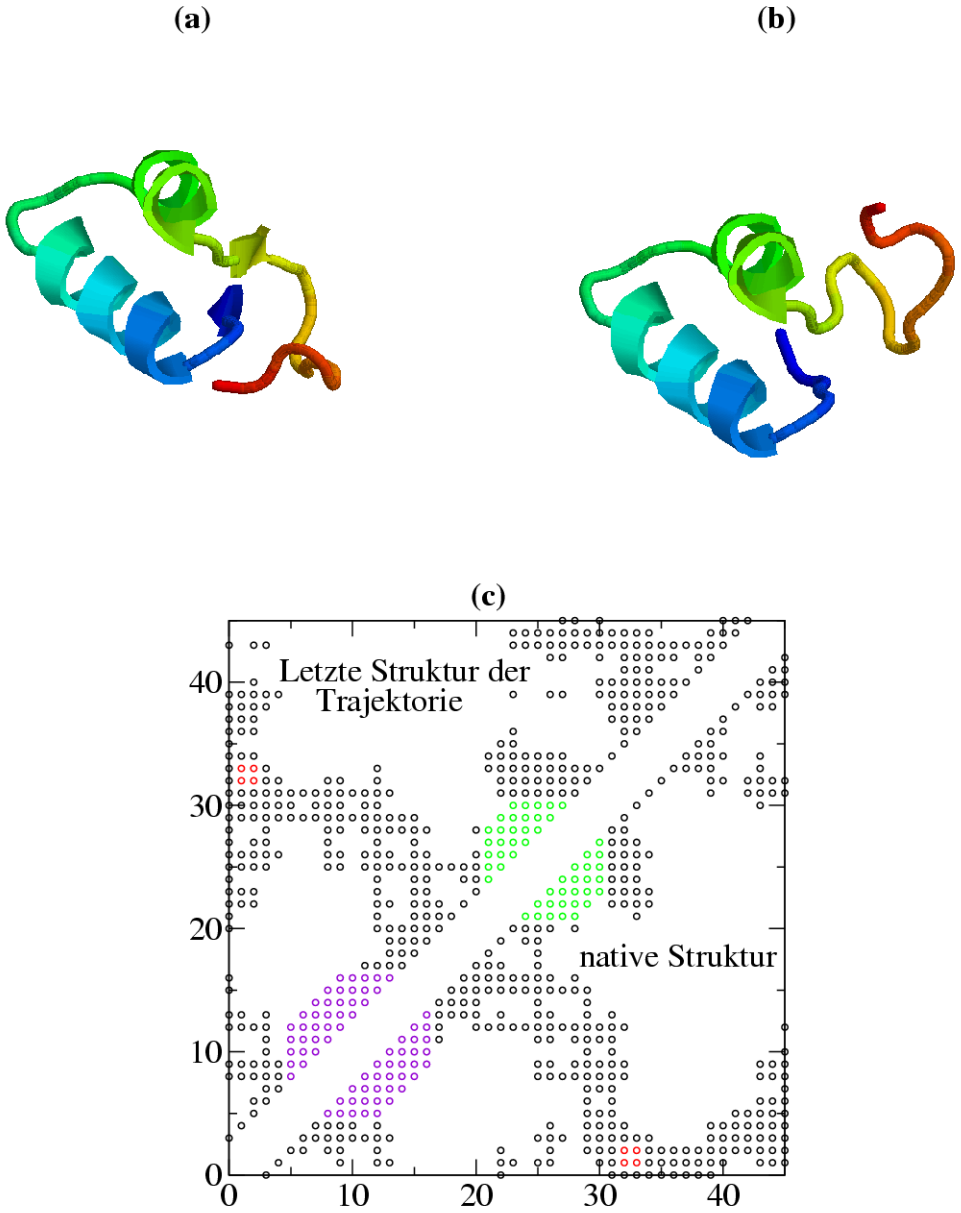

Abbildung 3.22: Struktur und Kontaktmatrix von Crambin vor und nach einer Monte Carlo Simulation mit 10^6 Monte Carlo Schritten im Raume der (ϕ, ψ)-Torsionswinkel bei 0K. (a): das native Crambin (b): Struktur nach der Simulation, (c): die zugehörigen Kontaktmatrizen. Die Kontaktenergieparameter sind mit Hilfe der linearen Optimierung nach Gleichung 2.24 mit einem Polynom in $(1-q)$ vom Grade $\Delta = 5$ (siehe 2.25) unter Verwendung des Sets_{420} erzeugt. Crambin ist in Set_{420} nicht enthalten. Zusätzlich zu den Kontaktenergieparametern werden die abstoßenden Potentiale nach Gleichung 2.9 und 2.10 verwendet.

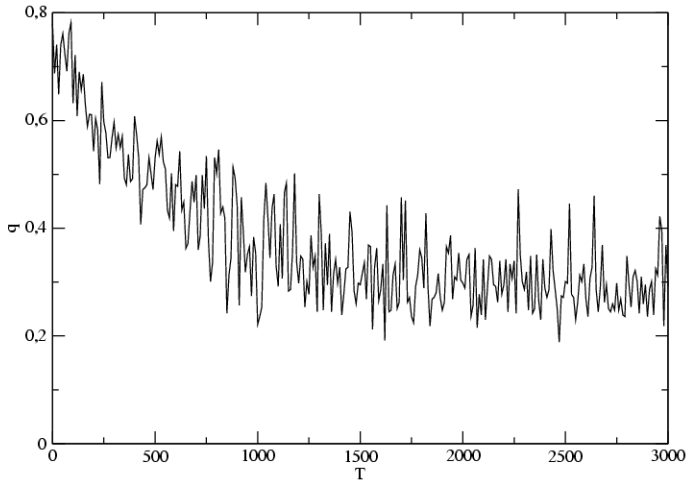

Abbildung 3.23: *Overlap* von nativem Crambin und Crambin nach 10^6 Monte Carlo Schritten im Raume der (ϕ, ψ)-Torsionswinkel in Abhängigkeit von der Temperatur. Die Kontaktenergieparameter sind mit Hilfe der linearen Optimierung unter Verwendung von Set_{420} erzeugt.

mulation $5 \cdot 10^7$ Monte Carlo Schritte. Am Ende der Simulation beträgt die Temperatur somit 13.14K. Wird die Energiefunktion mit Hilfe von Set_{135} und der nativen Struktur von Crambin trainiert, so ergibt sich ein *Overlap* von $q = 0.58$ und eine C_α cRMSD von 5.88Å. Für das Protein 2erl ergibt sich bei gleichen Bedingungen ein *Overlap* von $q = 0.70$ bei einer cRMSD von 7.27Å. Als einzige Zielsequenz dient hierbei 2erl. Als Trainingsset wird Set_{420} verwendet. Beim Protein *1orc* ergibt sich ein *Overlap* von $q = 0.55$ bei einer cRMSD von 6.99Å. Auch hier dient Set_{420} als Trainingsset, *1orc* wird als Zielsequenz verwendet.

Im Prinzip lässt sich der gleiche Test für die quasichemische Methode durchführen. In den meisten Fällen können mit einer einzelnen Sequenz nicht alle Kontakte realisiert werden. So können bei Crambin aufgrund der Sequenz z.B. nur 117 der 210 Kontakte auftreten. Von diesen 117 möglichen Paaren sind wiederum nur 96 in der nativen Struktur als Kontakt realisiert. Wird das Set_{45} zur Erzeugung der *Decoys* verwendet, so werden unter den *Decoys* Kontakte für alle 117 Paare realisiert. Für einen Kontakt i, der in der nativen Struktur nicht realisiert wird, der aber unter den *Decoys* mindestens einmal vorkommt, ergibt sich nach Gleichung 2.35 ein Kontaktenergieparameter $u_i = \infty$. Für die Unterscheidung zwischen nativen und nicht-nativen Strukturen ist dies eine gewünschte Eigenschaft der Energiefunktion: liegt ein solcher Kontakt in einer gegebenen Struktur vor, so ist klar, dass es sich nicht um die native Struktur handeln kann. Allerdings wird auf diese Weise allen Strukturen mit nicht-nativen Kontakten eine unendlich positive Energie zugeordnet. Sollen Strukturen, die der nativen Struktur

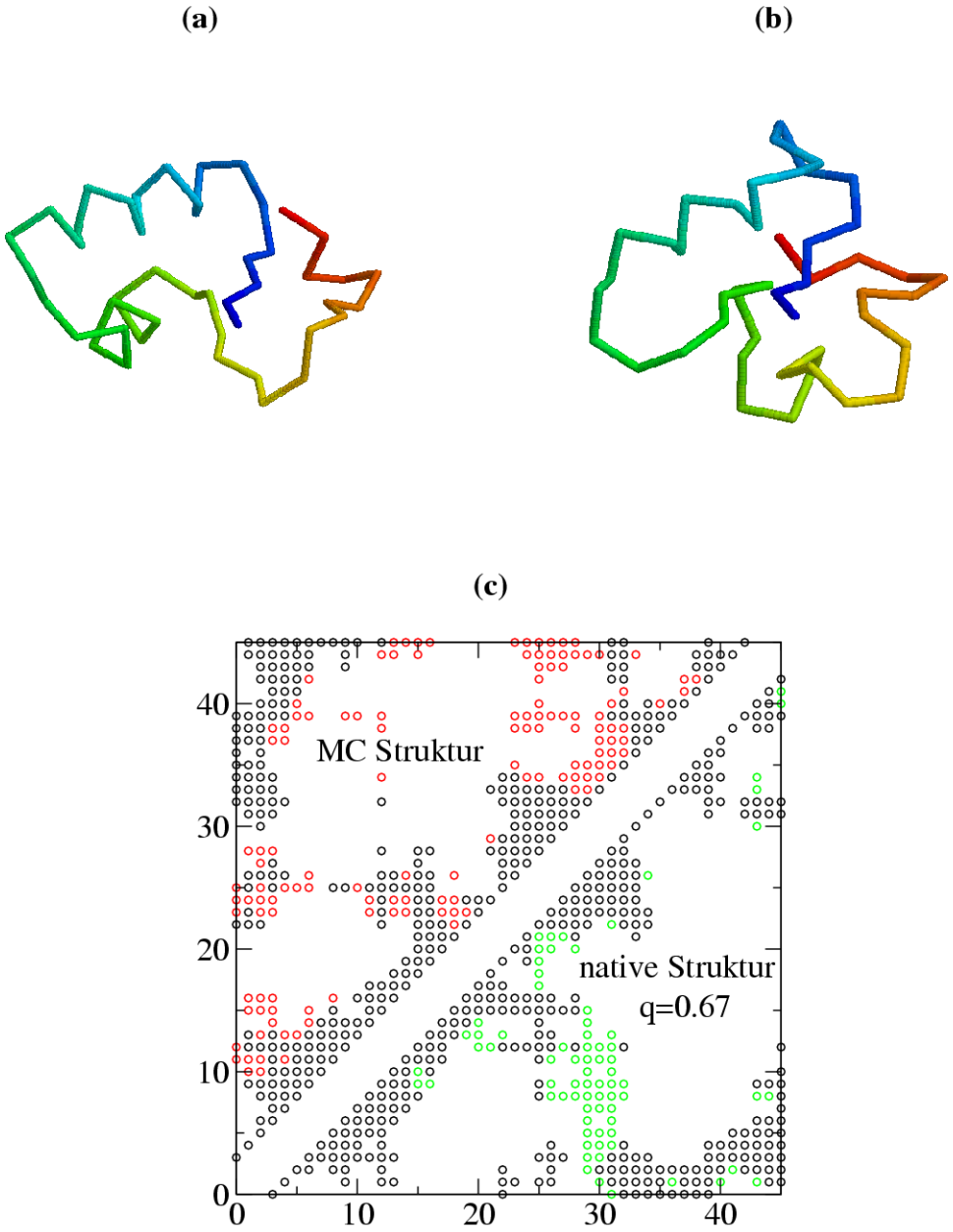

Abbildung 3.24: Struktur von Crambin nach eine MC Simulation, ausgehend von einer gestreckten Struktur (b) im Vergleich zu nativem Crambin (a) sowie die entsprechenden Kontaktmatrizen (c). Das Training der Energiefunktion erfolgt unter Verwendung der linearen Optimierung 2.24 unter Verwendung eines Polynoms in $(1-q)$ (siehe 2.25). Die nicht-nativen Strukturen für das Training sind durch *Threading* von Crambin durch Set_{420} erzeugt.

ähnlich sind, als ähnlich erkannt werden, so ist dies nicht sinnvoll.
Für die Monte Carlo Simulationen bedeutet ein unendlich positiver Energieparameter für einen nicht-nativen Kontakt, dass alle Strukturen abgelehnt werden die einen entsprechenden nicht-nativen Kontakt aufweisen. Solche Strukturen können jedoch der nativen Struktur sehr ähnlich sein, sind dann aber von der Energie her vollkommen nicht-nativ. Es bietet sich also an, entweder diese Kontakte als „neutrale Kontakte“ zu betrachten, ihnen also eine Energie von Null zuzuweisen, oder aber eine moderate positive Energie zu verwenden.
Tatsächlich ergibt sich mit einer moderaten Energie von +0.586[2] nach der Simulation ein *Overlap* von 0.72 und eine C_α cRMSD von 4.29Å. Abb. 3.25 zeigt die Kontaktmatrix dieser Struktur im Vergleich mit der von nativem Crambin.

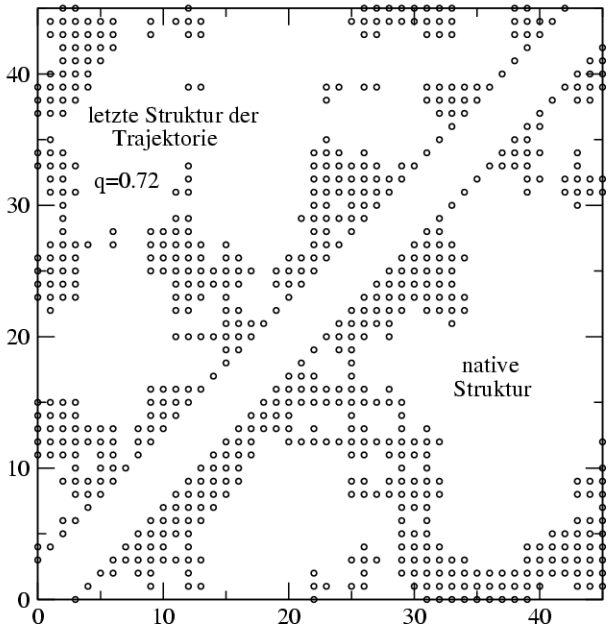

Abbildung 3.25: Die Kontaktmatrix von Crambin nach einer Monte Carlo Simulation im Raume der (ϕ,ψ)-Torsionswinkel sowie von nativem Crambin. Die Kontaktenergieparameter sind mit Hilfe der quasichemischen Methode erzeugt.

3.4.2 Erzeugung von *Decoys* mittels Monte Carlo

Strukturen aus Monte Carlo Simulationen lassen sich auch als *Decoys* zum trainieren von Energiefunktionen nutzen. Hierbei lässt sich die Ähnlichkeit der *Decoys* zu

[2] bei einer Skalierung der Parameter, so dass $\sum_{i=1}^{117} u_i^2 = 10$.

u_{nn}	0	0.257	0.432	0.586	0.685
q	0.64	0.66	0.69	0.72	0.63
C_α cRMSD	6.67Å	6.11Å	5.36Å	4.29Å	5.47Å

Tabelle 3.22: *Overlap* q und C_α cRMSD von Crambin nach $5 \cdot 10^7$ Monte Carlo Schritten. Das Training der Kontaktenergieparameter erfolgt nur für Crambin nach Gleichung 2.33. Die nicht-nativen Trainingsstrukturen sind mittels *Threading* von Crambin durch Set_{135} erzeugt. Kontaktenergieparameter für Kontakte, die in der nativen Struktur nicht auftreten weisen nach Gleichung 2.35 eine unendlich hohe Energie auf. Da es nicht sinnvoll ist, Strukturen mit solchen Kontakten in MC Simulationen grundsätzlich abzulehnen, werden verschiedene Werte für die entsprechenden nicht-nativen Kontaktenergieparameter u_{nn} ausprobiert.

den entsprechenden nativen Strukturen praktisch beliebig steuern. Sie hängt ab von der verwendeten Energiefunktion, der Temperatur und der Länge der Trajektorie. Ist die Energiefunktion in der Lage eine native Struktur zwischen beliebigen *Decoys* zu erkennen und ist sie gut korreliert, so wird sich bei Verwendung einer nativen Struktur als Startpunkt bei niedriger Temperatur die Struktur entlang der Trajektorie nicht stark verändern. Wird die Temperatur erhöht, so erfolgt eine Auffaltung. Durch Verwendung von mehreren kurzen Trajektorien bei niedrigen Temperaturen lässt sich eine große Zahl von Strukturen mit hoher Ähnlichkeit erzeugen. Es liegt also nahe, mit dieser Methode, zumindestens zusätzlich zu den mittels *Threading* erzeugten *Decoys*, weitere *Decoys* mit höherer Ähnlichkeit zu erzeugen.

Die Strukturen In Abb. 3.26 sind die Häufigkeitsverteilungen des *Overlaps* für Serien von Monte Carlo Simulationen dargestellt. Hierbei wird bei OK begonnen und die Temperatur in Schritten von 10K auf 200K erhöht. Dann wird in Schritten von 1K auf 400K erhöht und schließlich in Schritten von 40K auf 1000K. Bei der einen Simulation erfolgen in jedem Schritt 10^4 Monte Carlo Schritte mit der nativen Struktur von Crambin als Startpunkt, bei der anderen Simulation erfolgen pro Simulation $4 \cdot 10^5$ Monte Carlo Schritte. Hierbei wird insgesamt eine sehr große Zahl von Strukturen generiert. Besonders die ersten Schritte einer Trajektorie können Strukturen mit einem *Overlap* von $q = 1$ enthalten. Als *Decoys* verwendet werden jedoch nur Strukturen mit einem *Overlap* $q < 1$.
Die Zahl der erzeugten Strukturen hängt von der Akzeptanzrate ab. Im ersten Schritt bei 0K liegt sie für beide Serien bei 0.04, es werden also rund 400 bzw. 16000 Strukturen generiert. Bei 1000K liegt sie bei 0.61 für 10^4 Schritte bzw. 0.62 für $1 \cdot 10^5$ Schritte.

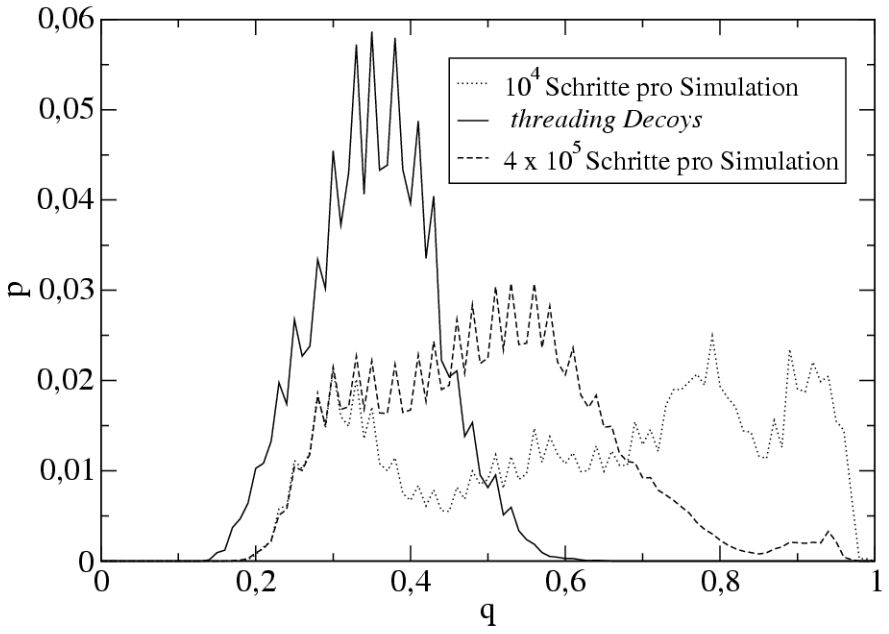

Abbildung 3.26: *Decoys* verschiedenster Ähnlichkeit lassen sich über Monte Carlo Simulationen erzeugen. Dargestellt sind die Häufigkeitsverteilungen von q nach Serien von Monte Carlo Simulationen bei verschiedenen Temperaturen sowie von q für *Decoys* erzeugt für Crambin mittels *Threading* unter Verwendung von Set_{135}. Die Monte Carlo Simulationen erfolgen mit der nativen Struktur von Crambin als Startpunkt. In einem Fall werden pro Temperaturschritt 10^4 Monte Carlo Schritte verwendet im anderen Fall $4 \cdot 10^5$ Monte Carlo Schritte.

Die Energiefunktion für die Simulationen ist mit Hilfe von *Threading* und der linearen Optimierung generiert worden. Als Zielsequenz diente hierbei Crambin, die *Decoys* wurden durch *Threading* dieser Sequenz durch das Set_{135} erzeugt[3]. Die Häufigkeitsverteilung dieser *Decoys* ist in Abb. 3.26 dargestellt.

Die Kontaktenergieparameter Die Berechnung der Parameter erfolgt mit der nativen Struktur von Crambin sowie allen *Decoys* aus den beiden Monte Carlo Serien. Die Korrelationen zwischen diesen Parametern und den Energieparametern aus der *Threading* Prozedur sind in Abb. 3.28 dargestellt. Obwohl die Optimierungen der Parameter mit jeweils zwei völlig verschiedenen *Decoy* Sets erfolgte, sind sich die Parameter relativ ähnlich. Für den großen Satz an Monte Carlo Strukturen ergibt sich ein Korrelationskoeffizient zu den *Threading* Parametern von $r = 0.776$, für den kleinen Satz

[3]Crambin hat eine Länge von 46 Aminosäuren. Das kleinste Protein in Set_{135} hat eine Länge von 60 Aminosäuren. Für die Erzeugung der *Decoys* wurden also alle 135 Proteine verwendet.

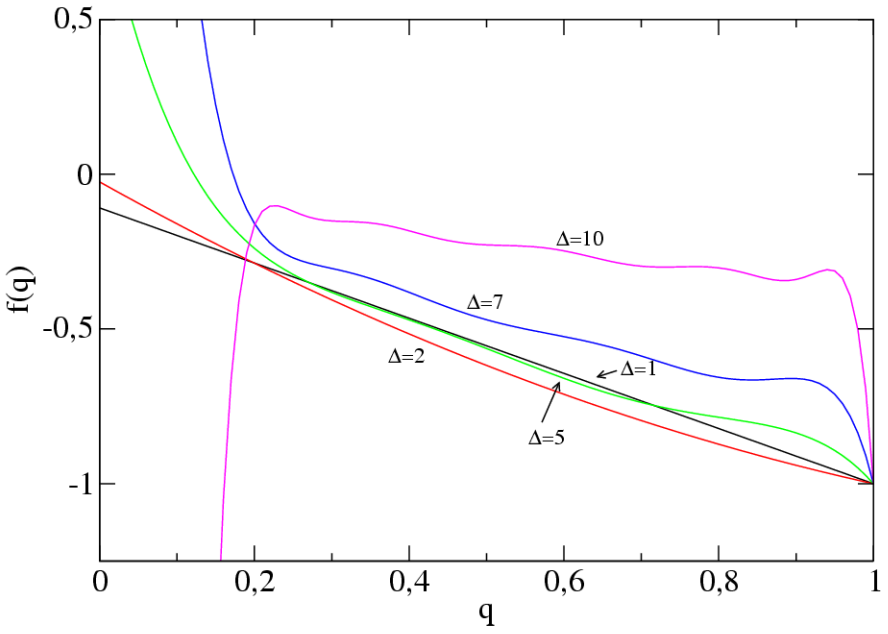

Abbildung 3.27: Mit Hilfe von Monte Carlo Simulationen ist es möglich ein breites Spektrum an *Decoy* Strukturen zu erzeugen. Werden die Simulationsbedingungen sinnvoll gewählt, so sind die Strukturen über einen weiten Ähnlichkeitsbereich zur nativen Struktur verteilt. Wird die Energiefunktion mittels der linearen Optimierung nach Gleichung 2.24 erzeugt, so unterscheiden sich die Polynome deutlich von denen mittels *Decoys* aus *Threading* erzeugten Energiefunktionen (siehe Abb. 3.10).

ergibt sich $r = 0.740$. Für die Erzeugung der Parameter über die Monte Carlo Strukturen wurde ein Polynom vom Grade 5 verwendet. Abb. 3.27 zeigt die entsprechenden Polynome von verschiedenem Grad δ. Diese unterscheiden sich grundlegend von den Polynomen aus den *Threading* Prozeduren (siehe Abb. 3.10). Die Ähnlichkeiten der mittels *Threading* erzeugten Strukturen liegen für fast alle *Decoys* im Bereich $q = 0.2$ bis $q = 0.6$. Über diesen Bereich sind die Polynome ab einem Grad von $\delta = 3$ annähernd konstant. Für die Monte Carlo Strukturen ist der gesamte Bereich $q \in [0.23, 0.98]$ mit einer großen Zahl an Strukturen abgedeckt. Es liegt kein *Decoy* mit einem *Overlap* $q < 0.18$ vor. Für kleine q nehmen die Polynome teilweise extrem hohe oder niedrige Werte an. Für $\delta = 9$ und $q = 0$ ergibt sich ein Wert von $f(q) = 61$, für $\delta = 10$ und $q = 0$ ergibt sich $f(q) = -145$. Im Bereich ab ca. $q = 0.22$ verhält sich das Polynom in sinnvoller Weise, d.h die Energie korreliert negativ mit dem *Overlap*.

Erkennung der nativen Struktur zwischen Monte Carlo Strukturen Die Frage ist, wie schwierig die Monte Carlo Strukturen für die Energiefunktion zu lernen sind,

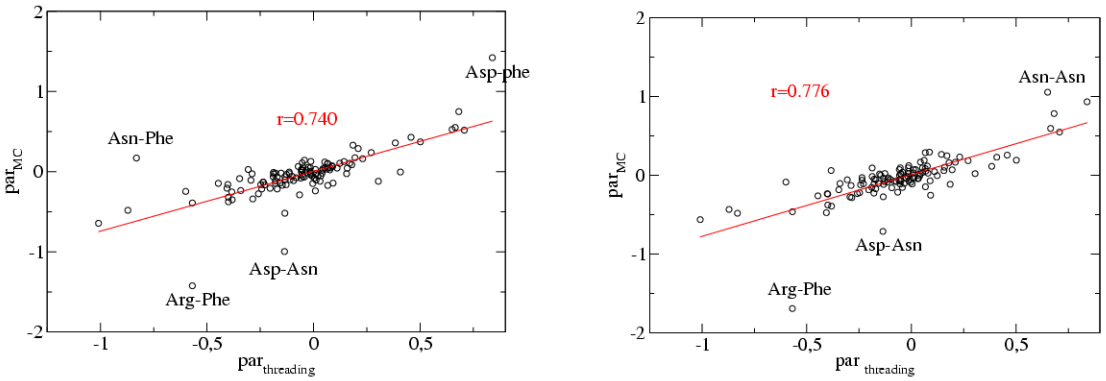

Abbildung 3.28: Die *Decoys* zur Erzeugung der Energieparameter aus Monte Carlo Simulationen unterscheiden sich in den Ähnlichkeiten entscheidend von denen aus *Threading*. Trotzdem ist die Korrelation der erzeugten Energieparameter relativ hoch. Links: Pro Temperaturschritt werden 10^4 MC-Schritte durchgeführt. Rechts: Pro Temperaturschritt werden $4 \cdot 10^5$ MC-Schritte durchgeführt. Hierbei wird eine sehr hohe Zahl an Strukturen erzeugt. Die Energieparameter aus *Threading* sind mit Set_{135}+Crambin erzeugt. Als einzige Zielsequenz dient hierbei Crambin.

also wie gut Strukturen mit nativen Eigenschaften aus dem Set an MC Strukturen erkannt werden. Die mittels des kleinen Satzes an Monte Carlo Strukturen trainierte Energiefunktion weist einer Struktur aus diesem Satz mit einem *Overlap* von $q = 0.95$ die niedrigste Energie (E=-28.79) zu. Für die native Struktur ergibt sich eine Energie von E=-27.90. Wird die mittels *Threading* Strukturen trainierte Energiefunktion verwendet, so weist die energieärmste Struktur (E=-42.63) einen *Overlap* von $q = 0.92$ auf. Die native Struktur hat hier eine Energie von E=-40.57. Beide Energiefunktionen erkennen also Strukturen die der nativen Struktur sehr ähnlich sind. Abb. 3.29 zeigt die Energien aller *Decoys* sowie der nativen Struktur unter Verwendung der beiden verschiedenen Energiefunktionen.

Auch die Korrelationen zwischen *Overlap* und Energie sind für beide Energiefunktionen sehr ähnlich. Trotz der hohen Zahl an *Decoys* (rund $3 \cdot 10^5$) ist schon die mittels *Threading* Strukturen trainierte Energiefunktion in der Lage die Energien sehr sinnvoll zuzuordnen. Wird die Zahl der *Decoys* stark erhöht und der große Satz an Monte Carlo Strukturen verwendet, so lässt die Erkennungsstärke dieser Energiefunktion nach. Der energieärmste *Decoy* aus diesem Satz hat einen *Overlap* zur nativen Struktur von $q = 0.71$ mit einer Energie von E=-44.19, die native Struktur weist eine Energie von E=-40.57 auf. Wird die Energiefunktion jedoch auf diese *Decoys* trainiert, so wird ein *Decoy* mit einer niedrigsten Energie von E=-31.99 und einem *Overlap* von $q = 0.93$ gefunden. Die native Energie beträgt hierbei E=-29.42. Es ist also möglich die Energiefunktion so auf diesen Satz an *Decoys* zu trainieren, dass ein *Decoy* aus dem nativen

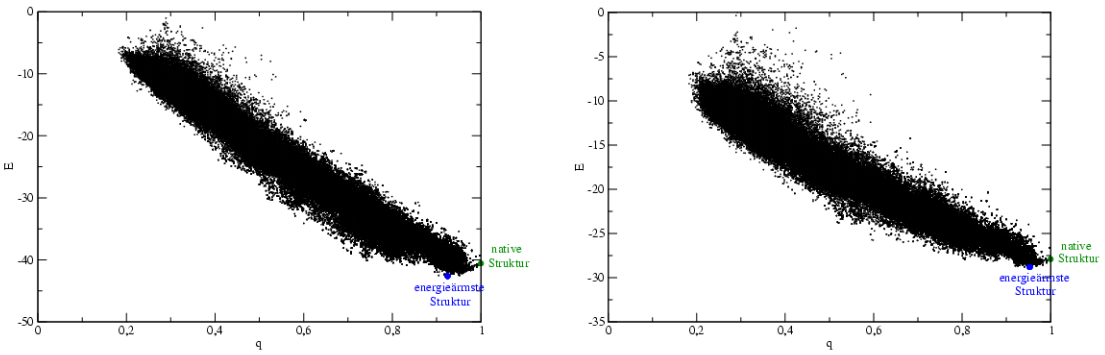

Abbildung 3.29: Mittels Monte Carlo Simulationen lässt sich eine hohe Zahl an *Decoys* zu einer gegebenen Sequenz erzeugen. Dargestellt sind die Energien solcher MC-Strukturen unter Verwendung verschiedener Energiefunktionen. Rechts: Die Energiefunktion ist mittels dieser Strukturen trainiert. Links: Die Energiefunktion ist trainiert mittels *Threading* Strukturen.

Bereich erkannt wird.

Faltungssimulationen mit den optimierten Parametern Werden die mittels Monte Carlo Strukturen optimierten Energieparameter für eine Faltungssimulation ausgehend von einer gestreckten Konformation verwendet, so ergibt sich eine Struktur mit einem *Overlap* von $q = 0.63$ und einer C_α cRMSD von 5.97Å zur nativen Struktur für die mittels des kleinen Satzes an MC-Strukturen optimierten Parameter. Die Parameter für die Monte Carlo Simulation entsprechen denen in 3.4.1.
Für die Kontaktenergieparameter aus dem großen Satz an MC-Strukturen ergibt sich ein *Overlap* von $q = 0.56$ und eine C_α cRMSD von 6.77Å. Es ergibt sich also keine Verbesserung gegenüber Parametern die nur über *Threading* von Crambin generiert sind.
Abb. 3.30 zeigt den Verlauf von Energie und *Overlap* dieser Monte Carlo Simulationen. Jeder 1000. angenommene Monte Carlo Schritt ist hier dargestellt.

Um ein noch größeres Spektrum an *Decoys* für das Training der Parameter zu erhalten, lassen sich die Monte Carlo *Decoys* mit den *Threading Decoys* kombinieren. Abb. 3.31 zeigt die *Overlap*-Verteilung für ein solches kombiniertes Set an *Decoys*. Eine *simulated annealing* Simulation mit den auf dieses Set angepassten Parametern ergibt am Ende der Trajektorie eine Struktur mit einem *Overlap* von $q = 0.63$ zu nativem Crambin. Die energieärmste Struktur weist einen *Overlap* von $q = 0.67$ und eine C_α cRMSD von 7.33Å auf.

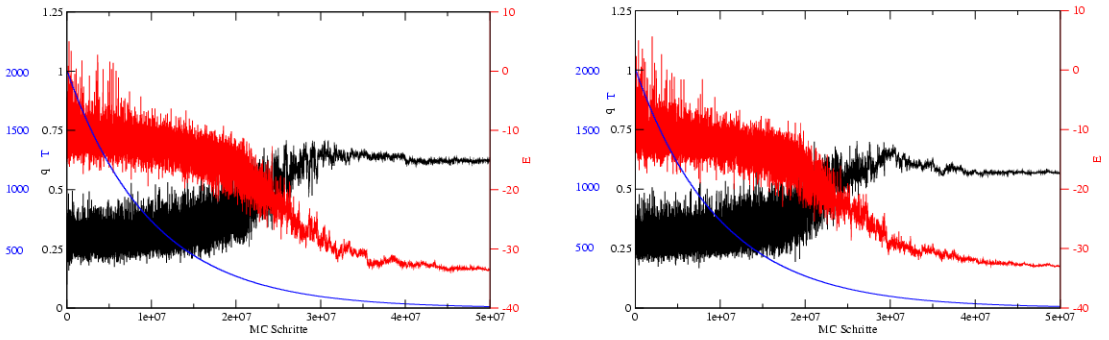

Abbildung 3.30: Verlauf von *Overlap*, Energie und Temperatur während einer Monte Carlo Simulation mit einer gestreckten Konformation von Crambin als Startpunkt. Dargestellt ist jeder 1000. angenommene MC Schritt. Die *Decoys* für das Training der Energiefunktion sind mit Hilfe von MC Simulationen von Crambin generiert. Oben: Für die Erzeugung der *Decoys* sind 10^4 MC Schritte je Temperaturschritt verwendet. Unten: Je Temperaturschritt erfolgen $4 \cdot 10^5$ MC Schritte.

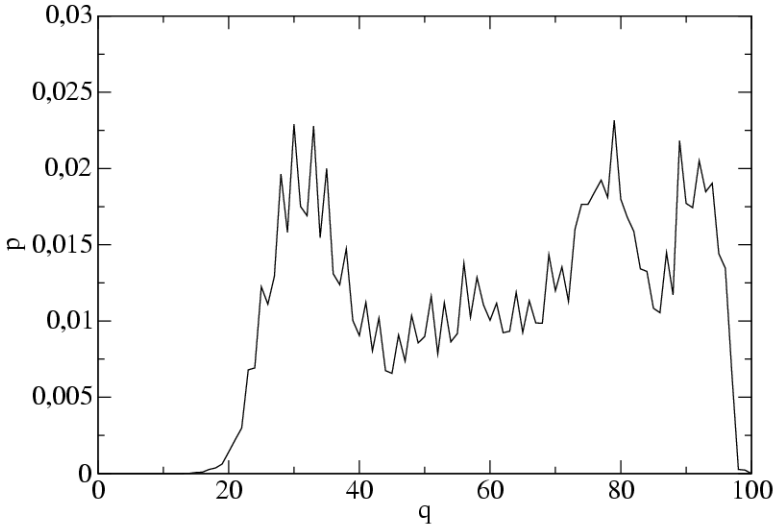

Abbildung 3.31: Mit einer Kombination aus *Threading* und Monte Carlo *Decoys* lässt sich ein breiter *Overlap* Bereich abdecken. Dargestellt ist die Verteilung des *Overlaps* für einen Satz aus Monte Carlo und *Threading* Strukturen. Das *Threading* erfolgt mit Hilfe von Crambin und Set_{135} mit Crambin als einziger Zielsequenz. Die Monte Carlo *Decoys* stammen aus dem kleinen Satz an MC Strukturen (10^4 MC Schritte je Temperaturschritt).

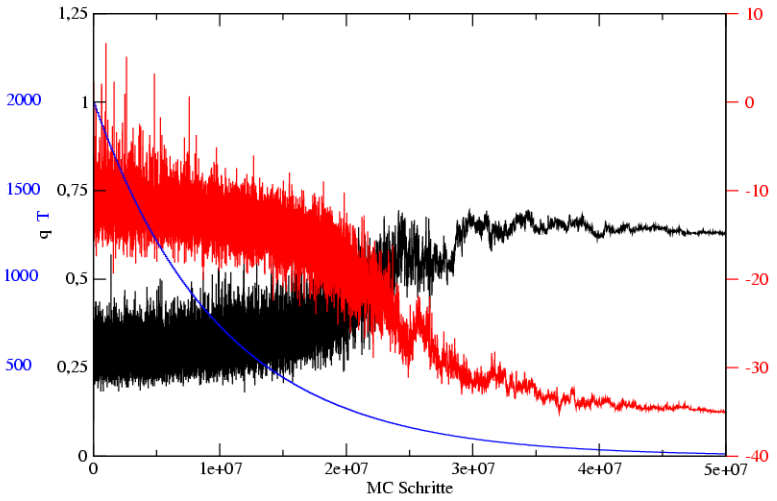

Abbildung 3.32: Verlauf von *Overlap*, Energie und Temperatur bei Monte Carlo Simulationen von Crambin. Die Energieparameter sind mit Hilfe eines kombinierten Sets an *Decoys* aus MC Simulationen und *Threading* optimiert.

Disulfidbrücken Auffällig sind die Energieparameter für Cystein-Cystein Paare. Diese betragen +0.05 für das kleine *Decoys* Set bzw. +0.09 für das große Set an *Decoys*. Crambin enthält sechs Cysteine, welche alle in Disulfidbrücken involviert sind. Trotz dieser starken Bindungen erhält man positive, also abstoßende Werte. Werden die Parameter mittels *Threading* anhand von Set_{135} optimiert, so ergibt sich für den Cystein-Cystein Parameter ein Wert von -1.34[4]. Das Vorliegen von drei Disulfidbrücken bedeutet natürlich, dass es $\binom{6}{2} \cdot \binom{4}{2} = 90$ Möglichkeiten gibt, diese aus sechs Cysteinen zu bilden. Der Cys-Cys Kontaktenergieparameter enthält keine Präferenz für bestimmte Cystein Paare. Möglicherweise enthält Crambin zuviele Disulfidbrücken bei sehr kurzer Kettenlänge.

3.4.3 Proteinstrukturvorhersage mit Hilfe von Monte Carlo Simulationen

Im vorangegangenen Kapitel wurde gezeigt, wie sich Faltungssimulationen mit Hilfe einer Monte Carlo Methode durchführen lassen. Wird als Startpunkt der Simulation

[4]Hierbei ist die Skalierung der Parameter nicht völlig identisch, da beim *Threading* mit Set_{135} N=210 Energieparameter vorliegen, hingegen bei Verwendung von Crambin als einzige Zielsequenz nur N=117. Für beide Sätze an Parametern gilt $\sum_{i=1}^{N} u_i^2 = 10$

eine gestreckte Konformation verwendet und sind die Energieparameter ohne Informationen der entsprechenden nativen Struktur erzeugt worden, so liegt eine echte Strukturvorhersage vor. Tabelle 3.23 zeigt für drei verschiedene Proteine die *Overlaps* und cRMSDs mit der nativen Struktur solcher Vorhersagen unter Verwendung verschiedener Energiefunktionen.

	Crambin		2erl		1orc	
verwendete Energiefunktion	q	cRMSD	q	cRMSD	q	cRMSD
LO_{45}	0.49	8.39	0.49	7.91	0.40	10.76
LO_{135}	0.55	9.71	0.50	7.97	0.44	12.11
LO_{420}	0.52	8.10	0.62	6.67	0.46	10.50
LO_{1014}	0.49	9.14	0.59	8.90	0.43	11.84
QCM_{45}	0.43	9.14	0.57	8.32	0.42	9.34
QCM_{135}	0.49	8.92	0.61	6.76	0.45	11.91
QCM_{420}	0.52	8.77	0.61	6.85	0.39	11.19
QCM_{1014}	0.55	8.46	0.57	8.92	0.42	12.33
$QCMw_{45}$	0.56	6.66	0.62	5.84	0.40	9.56
$QCMw_{135}$	0.52	6.50	0.62	8.87	0.42	11.26
$QCMw_{420}$	0.48	9.00	0.57	7.38	0.47	9.62
$QCMw_{1014}$	0.54	8.30	0.58	8.97	0.46	9.86
$boltz_{45}$	0.50	8.70				
$boltz_{135}$	0.50	8.64				
$boltz_{420}$	0.57	8.50				
$boltz_{1014}$	0.50	7.00				

Tabelle 3.23: *Overlaps* und C_α cRMSDs von Vorhersagen mittels einer Monte Carlo Simulation. Verschiedene Methoden und Proteinsets zur Erzeugung der Kontaktenergieparameter kommen zur Anwendung. Die nativen Strukturen der vorherzusagenden Proteine sind in den jeweiligen Trainingssets nicht enthalten.
LO_x: lineare Optimierung (Gleichung 2.24) unter Verwendung von Proteinset x
QCM_x: Quasichemische Methode (Gleichung 2.33) unter Verwendung von Proteinset x
$QCMw_x$: Quasichemische Methode mit Gewichtung (Gleichung 2.35) unter Verwendung von Proteinset x
$boltz_x$: Boltzmann-gewichtete Optimierung (siehe 2.9) unter Verwendung von Proteinset x.

Crambin (1ejg) Hinsichtlich des *Overlaps* ist für Crambin die Boltzmann-gewichtete Optimierung (siehe 2.9) am erfolgreichsten. Bezüglich der C_α cRMSD ist eine Energiefunktion, die mittels der quasichemischen Methode mit Gewichtung (siehe

Gleichung 2.35) erzeugt ist, am erfolgreichsten. Bei der Unterscheidung von nativen und nicht-nativen Strukturen (siehe Tabelle 3.13) zeichnet sich diese Funktion durch eine sehr gute Übertragbarkeit aus. Die Eigenschaft, Strukturen gut zu bewerten ohne vorher auf diese Strukturen gelernt worden zu sein, zeigt sich also auch bei der Strukturvorhersage.

Mating Pheromone Er-1 **(2erl)** Unter Verwendung einer Energiefunktionen, die mittels der quasichemischen Methode mit Gewichtung (siehe Gleichung 2.35) erzeugt ist, wird die niedrigste cRMSD (5.84Å) erreicht. Der höchste *Overlap* von $q = 0.62$ wird ebenfalls von dieser quasichemischen Methode sowie von der linearen Optimierung (siehe 2.10) erreicht.

Cro Repressor Insertion Mutant K56-[Dgevk] **(1orc)** Der höchste *Overlap* wird mit der quasichemischen Methode mit Gewichtung (Gleichung 2.35) erreicht, die niedrigste cRMSD mit der quasichemischen Methode nach Gleichung 2.33.

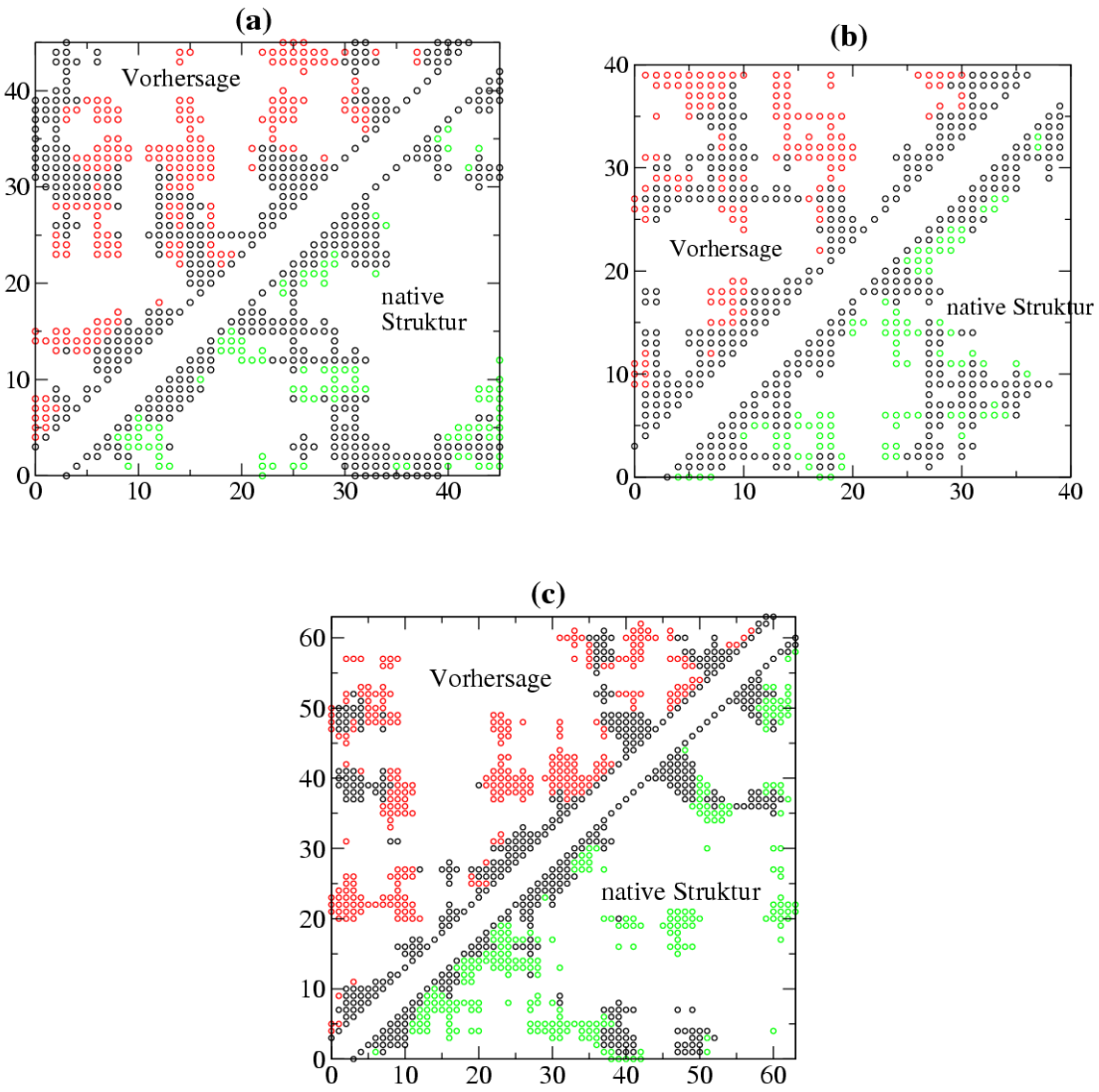

Abbildung 3.33: Kontaktmatrizen nach MC Simulationen ausgehend von gestreckten Konformationen. Die Erzeugung der Kontaktenergieparameter erfolgt jeweils mittels der quasichemischen Methode mit Gewichtung (siehe Gleichung 2.35)
(a): Crambin. Als Trainingsset dient Set_{45}. Crambin ist in diesem Set nicht enthalten.
(b): *2erl*. Als Trainingsset dient Set_{45}. *2erl* ist in diesem Set nicht enthalten.
(c): *1orc*. Als Trainingsset dient Set_{420} ohne *1orc*.