

# Proteinstrukturanalyse und -vorhersage mit einer optimierten Energiefunktion

Dissertation zur Erlangung des  
akademischen Grades des Doktors  
der Naturwissenschaften (Dr. rer. nat)

eingereicht im Fachbereich Biologie, Chemie,  
Pharmazie der Freien Universität Berlin

vorgelegt von  
**JOCHEN FARWER**  
aus Hamburg

Mai 2005

1. Gutachter: Prof. Dr. E. W. Knapp, Freie Universität Berlin

2. Gutachter: Prof. Dr. H. G. Holzhütter, Charité Berlin

Disputation am: 11.11.2005

# Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>1</b>
<b>2</b>	<b>Methoden</b>	<b>5</b>
2.1	Energiefunktionen für die Proteinfaltung . . . . .	5
2.2	Das Proteinmodell . . . . .	8
2.2.1	Die Kontaktmatrix . . . . .	8
2.2.2	Die Energiefunktion . . . . .	10
2.2.3	Disulfidbrücken . . . . .	11
2.2.4	$C_\alpha$ oder <i>all atom</i> Kontaktkriterium? . . . . .	11
2.2.5	Berechnen der Kontaktmatrizen . . . . .	12
2.3	Ähnlichkeits- und Distanzmaße . . . . .	13
2.3.1	cRMSD . . . . .	14
2.3.2	dRMSD . . . . .	14
2.3.3	<i>Power distance</i> . . . . .	14
2.3.4	<i>Overlap</i> . . . . .	15
2.4	Test der Distanzmaße . . . . .	15
2.4.1	Ein diskretes Proteinmodell . . . . .	16
2.4.2	<i>Build up</i> Algorithmus . . . . .	17
2.4.2.1	Optimierung der Winkelpaare . . . . .	17
2.4.2.2	Verwendete Proteine . . . . .	18
2.5	Monte Carlo Simulationen . . . . .	19
2.6	Erzeugen von <i>Decoys</i> . . . . .	20
2.6.1	Erzeugen von <i>Decoys</i> mittels <i>Threading</i> . . . . .	21
2.6.2	Erzeugen von <i>Decoys</i> mittels einer Monte Carlo Methode . . . . .	21
2.7	Verwendete Proteine und Strukturen . . . . .	21
2.8	Aminosäuren an der Proteinoberfläche . . . . .	26
2.9	Maximierung des Boltzmann-gewichteten <i>Overlaps</i> . . . . .	28
2.10	Abweichungen eines Gleichungssystems . . . . .	29

2.10.1	Wichtungsfaktoren . . . . .	33
2.10.2	Singulärwertzerlegung . . . . .	34
2.11	Eine quasichemische Methode . . . . .	35
2.12	Strukturvorhersage . . . . .	37
<b>3</b>	<b>Ergebnisse</b>	<b>41</b>
3.1	Eigenschaften von <i>Decoys</i> und nativen Proteinen . . . . .	41
3.1.1	Vergleich nativer Strukturen mit <i>Decoy</i> Strukturen aus der <i>Threading</i> Methode . . . . .	41
3.1.2	Packungsdichte in nativen Proteinen . . . . .	44
3.2	Proteinmodelle . . . . .	44
3.2.1	Proteinmodelle ohne abstoßendes Potential . . . . .	47
3.2.2	Proteinmodelle mit abstoßendem Potential . . . . .	50
3.3	Eigenschaften der Energiefunktionen . . . . .	52
3.3.1	Erkennung von nativen Proteinstrukturen . . . . .	52
3.3.1.1	Erzeugen der Energieparameter durch eine Boltzmann-gewichtete Optimierung . . . . .	53
3.3.1.2	Erzeugen der Kontaktenergieparameter mittels einer linearen Optimierung . . . . .	53
3.3.1.3	Erzeugen der Kontaktenergieparameter mittels einer quasichemischen Methode . . . . .	61
3.3.2	Der Kontaktabstand . . . . .	67
3.3.3	Verschiedene Sequenzabstandsbereiche . . . . .	68
3.3.4	Ein zusätzlicher Parameter für Aminosäuren an der Protein-oberfläche . . . . .	70
3.3.5	Erkennen von Strukturen mit nativen Eigenschaften . . . . .	73
3.4	Monte Carlo Simulationen . . . . .	79
3.4.1	Faltungssimulationen . . . . .	79
3.4.2	Erzeugung von <i>Decoys</i> mittels Monte Carlo . . . . .	83
3.4.3	Proteinstrukturvorhersage mit Hilfe von Monte Carlo Simula-tionen . . . . .	90
<b>4</b>	<b>Zusammenfassung und Ausblick</b>	<b>95</b>
<b>5</b>	<b>Abstract</b>	<b>99</b>
<b>6</b>	<b>Anhang</b>	<b>103</b>
6.1	Die Proteinsets . . . . .	103

<i>INHALTSVERZEICHNIS</i>	III
6.2 Die 20 natürlichen Aminosäuren . . . . .	111
6.3 Abkürzungsverzeichnis . . . . .	112
6.4 Publikationen . . . . .	112
6.5 Poster . . . . .	113
6.6 Vorträge . . . . .	114
6.7 Lebenslauf . . . . .	115
6.8 Erklärung . . . . .	116
<b>Literaturverzeichnis</b>	<b>117</b>



# Abbildungsverzeichnis

1.1	Amid-Ebene und Torsionswinkel . . . . .	2
2.1	Strukturdarstellung und Kontaktmatrizen von Glutaredoxin . . . . .	10
2.2	Das diskrete Proteinmodell verwendet nur die $C_\alpha$ Atome einer Proteinstruktur. Diese wird über die $(\alpha, \tau)$ Winkel beschrieben. . . . .	16
2.3	Verteilung der $(\alpha, \tau)$ Winkel der 202 Zielproteine aus Set <sub>1014</sub> . . . . .	18
2.4	Verteilungen der <i>Overlaps</i> $q$ für die fünf für <i>Threading</i> verwendeten Proteinsets unter Verwendung des $C_\alpha$ (a) sowie des <i>all atom</i> (b) Kontaktkriteriums. Für das $C_\alpha$ Kontaktkriterium werden für das Set <sub>NMR</sub> die Kontaktabstände $r_c = 8\text{\AA}$ und $r_c = 11\text{\AA}$ verglichen. . . . .	24
2.5	Zum Testen verschiedener Energiefunktionen in Monte Carlo Faltungssimulationen werden drei verschiedene Proteine verwendet. (a): Crambin (1ejg), (b): <i>Mating Pheromone Er-1</i> (2erl), (c): <i>Cro Repressor Insertion Mutant K56-[Dgev]</i> (1orc). . . . .	25
2.6	Mittlere Anzahl der $C_\alpha$ Kontakte pro Residuum der 20 verschiedenen Aminosäuren für die 202 Zielproteine aus Set <sub>1014</sub> . . . . .	27
2.7	Histogramm der Anzahl der Nachbarn individueller Residuen. . . . .	28
3.1	Wahrscheinlichkeitsverteilung der <i>Overlaps</i> $q$ aller <i>Decoys</i> sowie der <i>Decoys</i> mit größtem <i>Overlap</i> ( $q_{max}$ ) zur nativen Struktur für das Set <sub>1014</sub> . . . . .	42
3.2	Histogramm der mittleren Anzahl der Kontakte für alle nativen Zielsequenzen aus Set <sub>1014</sub> sowie für alle mit diesem Set erzeugten <i>Decoys</i> . . . . .	43
3.3	Histogramm der Anzahl der <i>Decoys</i> $N$ , die eine kompaktere Struktur als die native Struktur besitzen, für die 202 einzelnen Zielsequenzen aus Set <sub>1014</sub> . . . . .	43
3.4	Mittlere Anzahl der Kontakte $2 \cdot \frac{N_c}{N}$ in Abhängigkeit von der Kettenlänge $N$ für Kristallstrukturen (Set <sub>1014</sub> ) bzw. NMR-Strukturen (Set <sub>NMR</sub> ) unter Verwendung des $C_\alpha$ Kontaktkriteriums. . . . .	45
3.5	Abhängigkeit der erreichten cRMSD in Abhängigkeit von $N_{keep}$ . . . . .	48

3.6	$C_\alpha$ - $C_\alpha$ Abstandsverteilungen realer Proteinstrukturen und der mittels verschiedener Ähnlichkeitsmaße angepassten Modellstrukturen. . . .	49
3.7	$C_\alpha$ - $C_\alpha$ Abstandsverteilungen realer Proteinstrukturen und der mittels verschiedener Ähnlichkeitsmaße angepassten Modellstrukturen. . . .	51
3.8	Erkennung unter Verwendung einer Exponentialfunktion $-e^{(\beta \cdot q)}$ . . .	54
3.9	Schätzfunktion für die Energie von Proteinstrukturen in Abhängigkeit vom <i>Overlap</i> $q$ , $f(q) = -De^{(\beta \cdot q)}$ für verschiedene Werte $\beta$ . . . . .	55
3.10	Polynome in $(1 - q)$ mit verschiedenem Grad $\Delta$ für Set <sub>135</sub> sowie das Histogramm der <i>Overlaps</i> aller Strukturen die mit diesem Set erzeugt werden. . . . .	58
3.11	Die Energieparameter für Set <sub>135</sub> unter Verwendung der linearen Optimierung. . . . .	60
3.12	Werden die Gewichte nicht erkannter Sequenzen iterativ erhöht, so steigt die Erkennung an. Für neun Sequenzen werden in der vorliegenden Simulation in jedem Schritt die Gewichte erhöht ohne dass eine Erkennung erreicht wird. . . . .	62
3.13	Die Erkennung der nativen Proteine lässt sich durch das Einführen unterschiedlicher Wichtungsfaktoren für die verschiedenen Sequenzen verbessern. . . . .	63
3.14	Korrelation der Kontaktenergieparameter erzeugt mit Set <sub>45</sub> , bzw. mit Set <sub>1014</sub> unter Verwendung der quasichemischen Methode. . . . .	64
3.15	Die Kontaktenergieparameter berechnet mit Hilfe der quasichemischen Methode mit Gewichtung. Das Training erfolgt einmal mit Hilfe von Set <sub>45</sub> und einmal mit Hilfe von Set <sub>1014</sub> . . . . .	67
3.16	Die Kontaktenergieparameter bestimmt nach der quasichemischen Methode mit Gewichtung. Das Training erfolgt mit Hilfe von Set <sub>45</sub> . .	68
3.17	Abstandsverteilungen für ausgewählte Aminosäurepaare im Vergleich zur Abstandsverteilung aller Aminosäurepaare. . . . .	69
3.18	Die Energieparameter für Sequenzabstände von drei oder vier (Bereich 1) sowie für Sequenzabstände größer als vier (Bereich 2). . . . .	70
3.19	Erkennung von nativen Proteinen bei Hinzunahme von Wechselwirkungen der Proteinoberfläche mit Lösungsmittelmolekülen. . . . .	71
3.20	Die zwanzig Energieparameter für Kontakte mit der Umgebung sowie die mittleren Kontaktzahlen für die verschiedenen Aminosäuren für die 82 Zielsequenzen aus Set <sub>135</sub> , bzw. die 202 Zielsequenzen aus Set <sub>1014</sub> . . .	72



3.21	Korrelation zwischen dem <i>Overlap</i> des <i>Decoys</i> geringster Energie $q_{\min}$ und dem <i>Z-Score</i> für die 14 Sequenzen aus $\text{Set}_{1014}$ mit größtem $q_{\max}$ unter Verwendung verschiedener Lernsets. . . . .	75
3.22	Struktur von Crambin nach $10^6$ Monte Carlo Schritten im Raume der $(\phi, \psi)$ -Torsionswinkel bei 0K, des nativen Crambins sowie die zugehörigen Kontaktmatrizen. . . . .	80
3.23	<i>Overlap</i> von nativem Crambin und Crambin nach $10^6$ Monte Carlo Schritten im Raume der $(\phi, \psi)$ -Torsionswinkel in Abhängigkeit von der Temperatur. . . . .	81
3.24	Struktur von Crambin nach einer MC Simulation, ausgehend von einer gestreckten Struktur (b) im Vergleich zu nativem Crambin (a) sowie die entsprechenden Kontaktmatrizen (c). . . . .	82
3.25	Die Kontaktmatrix von Crambin nach einer Monte Carlo Simulation im Raume der $(\phi, \psi)$ -Torsionswinkel sowie von nativem Crambin. Die Kontaktenergieparameter sind mit Hilfe der quasichemischen Methode erzeugt. . . . .	83
3.26	Häufigkeitsverteilungen von $q$ für <i>Decoys</i> erzeugt in Serien von Monte Carlo Simulationen sowie Häufigkeitsverteilung von $q$ für <i>Decoys</i> erzeugt mittels <i>Threading</i> . . . . .	85
3.27	Polynome von verschiedenem Grad $\delta$ nach Optimierung der Kontaktenergieparameter mittels <i>Decoys</i> aus Monte Carlo Simulationen. . . . .	86
3.28	Korrelationen der Energieparameter erzeugt mittels Monte Carlo bzw. erzeugt mittels <i>Threading</i> . . . . .	87
3.29	Energien von Strukturen aus Monte Carlo Simulationen unter Verwendung verschiedener Energiefunktionen. . . . .	88
3.30	Verlauf von <i>Overlap</i> , Energie und Temperatur während Monte Carlo Simulationen mit einer gestreckten Konformation von Crambin als Startpunkt. . . . .	89
3.31	Verteilung des <i>Overlaps</i> für einen kombinierten Satz aus Monte Carlo und <i>Threading</i> Strukturen. . . . .	89
3.32	Verlauf von <i>Overlap</i> , Energie und Temperatur bei Monte Carlo Simulationen von Crambin. Die Energieparameter sind mit Hilfe eines kombinierten Sets an <i>Decoys</i> aus MC Simulationen und <i>Threading</i> optimiert. . . . .	90

3.33 Kontaktmatrizen nach MC Simulationen ausgehend von gestreckten Konformationen verschiedener Proteine. Die Erzeugung der Kontaktenergieparameter erfolgt jeweils mittels der quasichemischen Methode mit Gewichtung. . . . .	93
--	----

# Tabellenverzeichnis

3.1	Optimierte $(\alpha, \tau)$ Winkel für ein Modell ohne abstoßendes Potential. . .	47
3.2	Optimierte $(\alpha, \tau)$ Winkel für ein Modell mit abstoßendem Potential. . .	50
3.3	Erkennung von nativen Proteinen unter Verwendung der Boltzmann- gewichteten Optimierung. . . . .	53
3.4	Erkennung der nativen Proteinstrukturen unter Verwendung von $f(q) = 1 - q$ und $f(q) = -q$ . . . . .	54
3.5	Erkennung von nativen Proteinen unter Verwendung der Stufenfunktio- on 3.4 für die Optimierung der Kontaktenergieparameter. . . . .	56
3.6	Erkennung von nativen Proteinen unter Verwendung der linearen Op- timierung in Abhängigkeit vom verwendeten Grad $\Delta$ des Polynoms. . .	57
3.7	Differenz zwischen rechter und linker Seite des Gleichungssystems in Abhängigkeit vom Grad $\Delta$ des Polynoms. . . . .	58
3.8	Erkennung für verschiedene Kombinationen aus Trainings- und Testset bei Verwendung der linearen Optimierung. . . . .	59
3.9	Erkennung von nativen Proteinen unter Verwendung der linearen Op- timierung in Abhängigkeit vom verwendeten Grad $\Delta$ des Polynoms für das <i>all atom</i> Modell. . . . .	59
3.10	Werden die Gewichte nicht erkannter Sequenzen iterativ erhöht, so steigt die Erkennung an. Für neun Sequenzen werden in der vorliegen- den Simulation in jedem Schritt die Gewichte erhöht ohne dass eine Erkennung erreicht wird. . . . .	61
3.11	Erkennung der nativen Proteine. Die Energieparameter werden mit Hilfe der quasichemischen Methode erzeugt. . . . .	63
3.12	Erkennung der nativen Proteine. Alle <i>Decoys</i> mit einem <i>Overlap</i> $q >$ $q_{\text{thr}}$ werden beim Lernen der Energieparameter vernachlässigt. . . . .	65
3.13	Erkennung der nativen Proteine. Die Energieparameter wurden mit Hilfe der quasichemischen Methode mit Gewichtung bestimmt. . . . .	65

3.14	Erkennung der nativen Proteine. Alle <i>Decoys</i> mit einem <i>Overlap</i> $q > q_{\text{thr}}$ werden beim Lernen der Energieparameter vernachlässigt. . . . .	66
3.15	Erkennung von nativen Proteine unter Verwendung des <i>all atom</i> Modells. Die Energieparameter werden mit Hilfe der quasichemischen Methode ohne Gewichtung bzw. mit Gewichtung erzeugt. . . . .	67
3.16	Erkennung von nativen Proteinen in Abhängigkeit vom Kontaktabstand $r_c$ . . . . .	69
3.17	Erkennung bei Verwendung von verschiedenen Sequenzabstandsbereichen. . . . .	71
3.18	Größte <i>Overlaps</i> die mit dem $\text{Set}_{1014}$ mittels <i>Threading</i> generiert werden sowie <i>Overlaps</i> und <i>Z-Scores</i> der jeweils energieärmsten <i>Decoys</i> für verschiedene Sequenzen. . . . .	76
3.19	Korrelationskoeffizienten $r$ zwischen dem <i>Overlap</i> des <i>Decoys</i> geringster Energie $q_{\text{min}}$ und dem <i>Z-Score</i> für die 14 Sequenzen aus $\text{Set}_{1014}$ mit größtem $q_{\text{max}}$ . . . . .	77
3.20	Ähnlichkeiten zwischen Zielsequenzen und Sequenzen die für die Erzeugung des <i>Decoys</i> mit größtem <i>Overlap</i> $q_{\text{max}}$ verwendet werden sowie die maximal vorhandene Ähnlichkeit unter allen verwendeten <i>Decoys</i> . . . . .	77
3.21	Größter <i>Overlap</i> $q_{\text{max}}$ aller <i>Decoys</i> sowie <i>Overlap</i> $q_{\text{min}}$ des energieärmsten <i>Decoys</i> für verschiedene Sequenzen. . . . .	78
3.22	<i>Overlap</i> $q$ und $C_\alpha$ cRMSD von Crambin nach $5 \cdot 10^7$ Monte Carlo Schritten. Das Training der Kontaktenergieparameter erfolgt nur für Crambin nach der quasichemischen Methode ohne Gewichtung. Es werden verschiedene Werte für die nicht-nativen Kontaktenergieparameter $u_{mn}$ ausprobiert. . . . .	84
3.23	<i>Overlaps</i> und $C_\alpha$ 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. . . . .	91
6.1	$\text{Set}_{\text{disk}}$ . . . . .	106
6.2	$\text{Set}_{\text{disk,keep}}$ . . . . .	106
6.3	$\text{Set}_{45}$ . . . . .	106
6.4	$\text{Set}_{135}$ ohne $\text{Set}_{45}$ . . . . .	107
6.5	$\text{Set}_{420}$ ohne $\text{Set}_{135}$ . . . . .	108

*TABELLENVERZEICHNIS*

XI

6.6	Set <sub>1014</sub> ohne Set <sub>420</sub> . . . . .	110
6.7	Liste der Aminosäuren mit Abkürzungen und Seitenketten . . . . .	111

