Studienarbeit: Faltungssimulationen für Gitterproteine

Daniel Maticzka

Betreuer: Martin Mann
Institut für Bioinformatik
Albert-Ludwigs-Universität Freiburg
7. April 2008

1 Einleitung

Gitterproteine sind vereinfachte Computermodelle von Proteinen, die dazu dienen, den Faltungsprozess von Proteinen zu erforschen. Über eine Energiefunktion kann jeder von einem Gitterprotein einnehmbaren Struktur eine Energie zu gewiesen werden. Zusammen mit dieser Energiefunktion bildet die Menge aller Strukturen die Zustände einer Energielandschaft. Über eine Transformationsfunktion können die Zustände der Energielandschaft ineinander überführt werden. Auf diese Weise kann innerhalb einer Energielandschaft nach Energieminima gesucht werden.
Um diese Faltungsimulationen durchführen zu können musste zunächst der in [1] beschriebene Satz von Transformationsfunktionen, die Pull Moves, von zwei auf dreidimensionale Gitter erweitert und implementiert werden. Aufbauend auf den BIU und ELL Bibliotheken [6] wurde daraufhin ein Kommandozeilentool, HPfold, erstellt, mit dem Faltungssimulationen mit Gitterproteinen im HP-Modell unter Verwendung von Pull Moves und einer Metropolis Monte Carlo Suche durchgeführt werden konnten. Zusätzlich mussten Werte für zwei Parameter, die grundlegenden Einfluss auf die Erfolgsquote der Suche haben, festgelegt werden.
Ziel war es nun, eine Klassifizierung von Gitterproteinen nach ihren Faltungseigenschaften vorzunehmen zu können. Dazu musste eine Reihe von Faltungssimulationen für eine Anzahl proteinähnlicher Gitterproteine durchgeführt werden. Proteinähnlich bedeutet, dass genau eine Struktur existiert, deren Energie niedriger ist als die aller anderen Strukturen. Für jedes dieser

Gitterproteine waren Energie und Konformation der funktionalen Struktur bereits über Constraint-basierte Methoden berechnet. Für 1.000 dieser Proteinsequenzen wurden jeweils 1.000 Durchläufe der Suche gestartet und die Erfolgsquote für das Auffinden der Struktur minimaler Energie festgehalten. Zum Abschluss folgt eine Auswertung dieser Daten mit Hinblick auf die Verwendung der Erfolgsquote der Suche zur Klassifikation der Faltungseigenschaften eines Gitterproteins.

2 Grundlagen

2.1 Proteinfaltung

Proteine sind essentieller Bestandteil von Organismen, wo sie unter anderem strukturelle und funktionale Aufgaben übernehmen. Proteine bestehen aus einer Kette von Aminosäuren, die aufgrund molekularer Bindungskräfte eine dreidimensionale Struktur annimmt. Diese Struktur bestimmt maßgeblich die Funktion eines Proteins. Die Abfolge der Aminosäuren wird durch ein Gen codiert.[4] Durch die Sequenzierung vollständiger Genome existieren große Datenmengen über Aminosäuresequenzen von Proteinen. Das Problem der Proteinfaltung bezeichnet die Vorhersage der Struktur eines Proteins bei bekannter Aminosäuresequenz.[5]
Detaillierte Proteinmodelle, bei denen teilweise einzelne Atome simuliert werden, sind extrem berechnungsintensiv. Mit heutiger Technologie können daher nur einige Mikrosekunden eines Faltungsvorgangs berechnet werden. Aus diesem Grund wird bei der Erforschung der Proteinfaltung auf stark vereinfachte Proteinmodelle zurückgegriffen, die Gitterproteine. Die Berechnung ist hier einfach genug, um vollständige Faltungsvorgänge durchführen zu können. Obwohl das Modell sehr einfach ist, bleiben die Haupteigenschaften der Problemstellung erhalten. Insbesondere ist das Faltungsproblem im HP-Modell sowohl für zweidimensionale als auch für dreidimensionale Gitter NP-schwer [3].

2.2 Gitterproteine

Ein Gitterprotein besteht aus einer Sequenz von Monomeren, die derart auf einem Gitter angeordnet werden, dass keine Position des Gitters mehrfach besetzt ist. (selfavoiding Walk) Dabei wird jede Aminosäure durch jeweils genau ein Monomer repräsentiert.[2] Über eine Energiefunktion wird jeder Struktur eine Energie zugewiesen. Im einfachen HP-Modell wird jede Aminosäure entweder als hydrophil (H) oder als hydrophob (P) markiert. Für

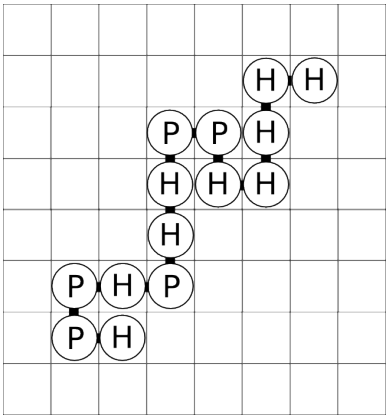

Abbildung 1: Gitterprotein auf zweidimensionalem quadratischen Gitter. Nach HP-Modell würde diesem eine Energie von -2 zugewiesen.

jeweils zwei hydrophile Aminosäuren wird der Struktur eine Energie von -1 zugewiesen, wenn diese benachbarte Stellen auf dem Gitter, aber nicht in der Sequenz einnehmen. Es wird davon ausgegangen, dass diejenige Struktur mit der niedrigsten Energie die funktionale Form, den "native state", des Proteins darstellt. Zusammen mit der Energiefunktion stellen die Strukturen eine Energielandschaft dar, in der nach der optimalen Konformation mit minimaler Energie gesucht werden kann. Zwei solcher Zustände sind benachbart, wenn sie durch eine Transformationsfunktion ineinander überführbar sind.

2.3 Pull Moves

Als Transformationsoperation für die Transformation zwischen zwei Zuständen wurden die in [1] beschriebenen Pull Moves verwendet. Das Pull Move Set ist eine Menge lokaler und vollständiger Transformationen. Aus der Vollständigkeit folgt, dass jede gültige Konformation durch eine Reihe von Pull Moves in jede andere gültige Konformation überführt werden kann. Dies ist eine wichtige Eigenschaft für die Suche in der Energielandschaft, da die optimale Lösung nicht gefunden werden kann, wenn die Transformationen nicht in der Lage sind, diesen Zustand zu generieren.
Sei in der folgenden Beschreibung v_i derjenige Knoten, der zuerst bewegt wird. Sei v_{i+1} derjenige Knoten, dessen Position nicht geändert wird. Von den Positionen von v_i und v_{i+1} ausgehend können nun die Positionen der Punkte C und L gewählt werden. Bei Durchführung der Transfomation wird zunächst der Knoten v_i auf Position L verschoben, dann wird der Knoten v_{i-1} auf Position C verschoben. (Abb. 2.) Dabei müssen jeweils die Positionen von v_{i+1} und L, von v_i und C, sowie von C und L auf dem Gitter benachbart sein.

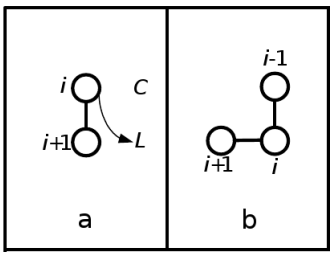

Abbildung 2: a) Ausgangszustand. b) Endzustand: Knoten v_i wurde auf Position L, Knoten v_{i-1} wurde auf Position C verschoben. Entnommen aus [1]

Um die Transformation abzuschließen, werden nun nacheinander die Knoten v_{i-2} bis v_0, beginnend bei Knoten v_{i-2}, nachgezogen. Dieser Vorgang bricht ab, sobald alle Knoten des Gitterproteins wieder eine zusammenhängende Kette ergeben. Daraus folgt direkt die Lokalität der Transformationen, da bei jeder Transformation möglichst wenige Knoten verschoben werden und kein Knoten weit von seiner ursprünglichen Position wegbewegt wird.

Im folgenden habe ich den Pull Move Algorithmus auf dreidimensionale Gitter erweitert und in die BIU-Bibliothek integriert. Diese Implementation nutzt die in der BIU vorhandenen Gitterbeschreibungen und kann somit auf zweidimensionale und dreidimensionale kubische Gitter (SQR, CUB), sowie das kubisch flächenzentrierte Gitter (FCC - face centered cubic) angewendet werden.

3 Tool und Parameteranpassung

3.1 Fragestellung

Gegeben war eine Reihe von Sequenzen minimaler freier Energie der Länge 27 im HP-Modell auf dreidimensionalem kubischem Gitter. Für jede dieser Sequenzen existiert jeweils nur eine optimale Konformation. Ziel war es nun, herauszufinden wie häufig diese optimale Struktur durch eine Metropolis Monte Carlo Suche auf der Energielandschaft unter Verwendung der Pull Moves gefunden werden kann, um basierend auf diesen Daten eine Einteilung der Sequenzen in solche mit guten und solche mit schlechten Faltungseigenschaften vornehmen zu können.

Für die Durchführung der Metropolis Monte Carlo Suche mussten zunächst Werte für die zwei offene Parameter gefunden werden: die Temperatur T der Metropolis Funktion, sowie die Anzahl der anzunehmenden suboptimalen Zustände, nach der die Suche erfolglos abbricht. Beide Parameter haben

beträchtliche Auswirkungen auf die Erfolgswahrscheinlichkeit und Dauer der Suche.

3.2 Metropolis Monte Carlo Suche

Bei der Metropolis Monte Carlo Suche werden aus einem Ausgangszustand Nachfolgezustände generiert, die jeweils mit einer gewissen Wahrscheinlichkeit angenommen werden. Im Fall der Annahme wird die Suche mit dem akzeptierten Zustand fortgesetzt, bis ein vordefiniertes Abbruchkriterium zutrifft. Die hier durchgeführte Suche wird nach Erreichen der optimalen Energie oder Überschreiten einer maximalen Anzahl akzeptierter Zustände abgebrochen.
Für die Bestimmung, ob ein Nachfolgezustand akzeptiert wird, wurde das nichtdeterministische Metropolis Kriterium verwendet. Die Transformation von Zustand S_1 mit Energie E_1 zu Zustand S_2 mit Energie E_2 wird akzeptiert, falls gilt

$$Rnd < e^{\frac{(E_1 - E_2)}{kT}},$$

wobei Rnd eine Zufallszahl zwischen 0 und 1 und k=1 ist. Dies hat zur Folge, dass Wechsel zu Zuständen mit gleicher oder niedriger Energie immer durchgeführt, Wechsel zu Zuständen mit höherer Energie jedoch nur mit Wahrscheinlichkeit $e^{\frac{(E_1 - E_2)}{kT}}$ durchgeführt werden.

3.3 Anpassung der Parameter

Daraus folgt, dass ein Folgezustand garantiert angenommen wird, wenn seine Energie niedriger als die des Ausganszustands ist. Ist die Energie des Folgezustands dagegen höher als die des Ausgangszustands, wird er nur mit einer gewissen Wahrscheinlichkeit angenommen, die mit steigender Energiedifferenz abfällt. Ist die Temperatur zu hoch gewählt, wird der Energieverlauf der angenommenen Zustände sehr sprunghaft und zufällig. Ist die Temperatur zu niedrig gewählt, werden Wechsel zu Zuständen mit höherer Energie mit sehr geringer Wahrscheinlichkeit akzeptiert. Dadurch wächst die Gefahr, dass die Suche in einem lokalen Energieminimum endet und sich nicht mehr davon lösen kann.
Um den Bereich für die Temperatur besser einschätzen zu können, habe ich für einige Sequenzen Suchläufe mit verschiedenen Temperaturwerten durchgeführt und die Energieverläufe während der Suche auftragen lassen. Nach diesen Ergebnissen sollte die Temperatur in einem Bereich zwischen $\frac{1}{6}$ und $\frac{1}{2}$ gewählt werden. Es ist zu erwarten, dass die Suche in einem begrenzten

Temperaturbereich gute Ergebnisse liefert, die Performanz aber an dessen Rändern abfällt, da die Suche dort zu chaotisch oder zu geradlinig verläuft. Abb. 3 zeigt einige Energieverläufe bei verschiedenen Temperaturen.

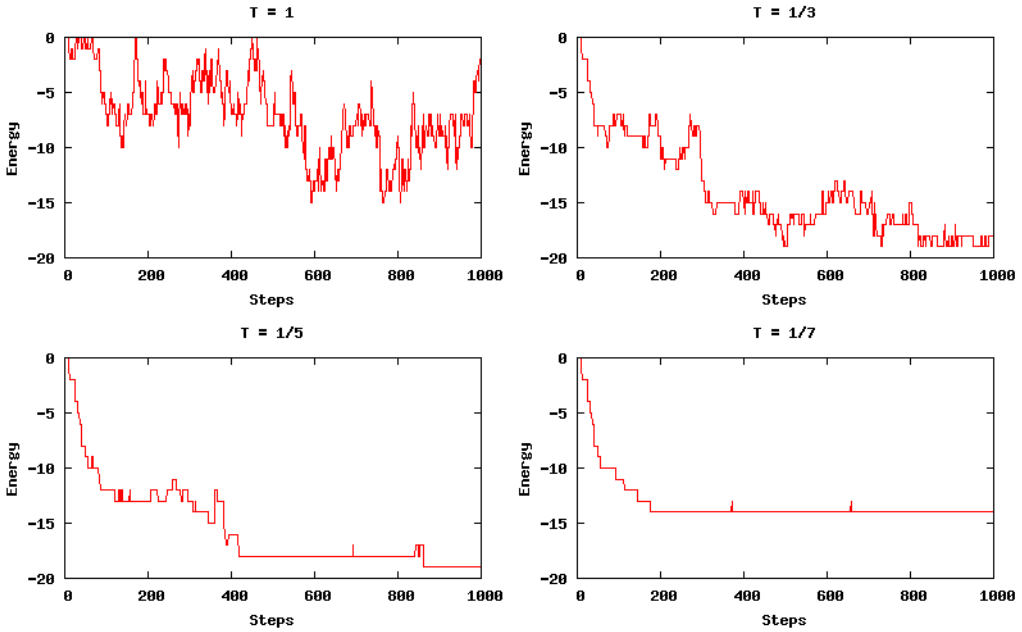

Abbildung 3: Energieverläufe für Suche mit Sequenz "HHHHHHHHPHPP-HHHHPHPHHPHPHPH" und Temperaturen 1, $\frac{1}{3}$, $\frac{1}{5}$ und $\frac{1}{7}$. Der Energieverlauf bei T=1 weist starke, zufällig erscheinende, Sprünge auf. Es ist keine Bewegung in Richtung optimaler Energie zu erkennen. Der Energieverlauf bei T=$\frac{1}{7}$ weist kaum Sprünge zu suboptimalen Zuständen auf. Die Suche endet in einem lokalen Minimum und kann sich nicht mehr daraus lösen. Die Energieverläufe für T=$\frac{1}{3}$ und T=$\frac{1}{5}$ sind unterschiedlich sprunghaft, erreichen aber beide eine Energie von -19.

Um den Zusammenhang zwischen Temperatur, maximaler Schrittzahl und Trefferwahrscheinlichkeit aufzudecken, habe ich für 11 verschiedene Sequenzen jeweils 1.000 Suchläufe mit maximal 10.000 Schritten für die Temperaturen in diesem Bereich durchgeführt und die Anzahl der Treffer aufsummiert. Durch diese Daten konnte die ursprüngliche Vermutung, dass die Performance jenseits eines bestimmten Bereichs abfällt, bestätigt werden. Abb. 4 verdeutlicht diesen Zusammenhang.
Außerdem konnte die Vermutung, dass einige Sequenzen bessere Faltungseigenschaften besitzen als andere, bestätigt werden. Für eine der elf Sequenzen wurde die optimale Konformation bei keiner Temperatur gefunden. Für drei weitere Sequenzen ist der Temperaturbereich, in dem Treffer erzielt werden

konnten, sehr klein. Die restlichen sieben Sequenzen zeigen einen konstanten Anstieg und anschließenden Abstieg der Performance in einem größeren Temperaturbereich. Die Trefferquote liegt bei diesen Sequenzen zwischen 1% und 7%.

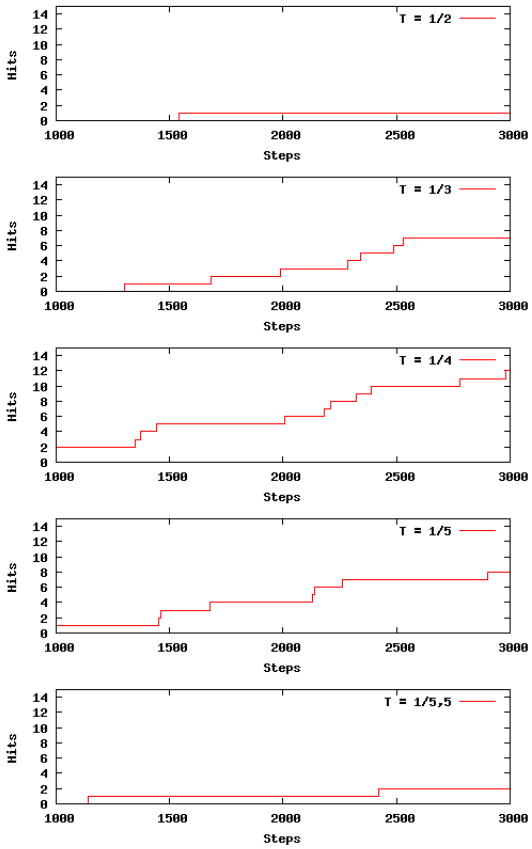

Abbildung 4: Ausschnitt der Auswertung von 1.000 Suchläufen mit jeweils maximal 10.000 Schritten für die Sequenz "HPHPHHHPHHPHHPHHPH-PHPHHPPHH". Aufgetragen ist die Summe der Treffer der optimalen Struktur im Bereich [0, Steps]. Die Performanz erreicht bei einer Temperatur von $\frac{1}{4}$ das Optimum und fällt in Richtung niedriger und höherer Temperaturen.

Um zu einer Abschätzung des Temperaturparameters für die folgende Untersuchung zu kommen, habe ich nun die durchschnittliche Trefferquote aller Sequenzen für jede Temperatur berechnet. Diese war maximal bei einer Temperatur von $\frac{1}{3{,}5}$, war jedoch nur geringfügig kleiner für T=$\frac{1}{3}$. Da die Suche mit T=$\frac{1}{3{,}5}$ im Schnitt um einen Faktor 1,7 langsamer war als die Suche mit T=$\frac{1}{3}$, habe ich mich entschieden, den Parameter T=$\frac{1}{3}$ zu wählen. Für die maximale

Anzahl Schritte erschien 4.000 als guter Kompromiss zwischen durchschnittlicher Trefferquote und Dauer der Suche.

4 Auswertung

Nun wurden für 1.000 der Sequenzen jeweils 1.000 Suchläufe bei Temperatur $\frac{1}{3}$ und maximal 4.000 Schritten gestartet und für jede Sequenz die Anzahl der Treffer der mfe-Struktur aufgetragen (Abb. 5). Für 95 Sequenzen wurde die optimale Struktur nie gefunden. Für 137 Sequenzen wurde die optimale Struktur einmal gefunden. Für einen bis 23 Treffer sinkt die Zahl der Sequenzen relativ stetig bis auf 1. Der Wert für 11 Treffer liegt mit 12 Sequenzen etwas abseits von diesem Trend. Zusätzlich zeigen 5 Sequenzen besonders gute Faltungseigenschaften mit 29, 29, 30, 32 und 55 Treffern.

Die Zahl der Sequenzen mit besonders guten Faltungseigenschaften ent-

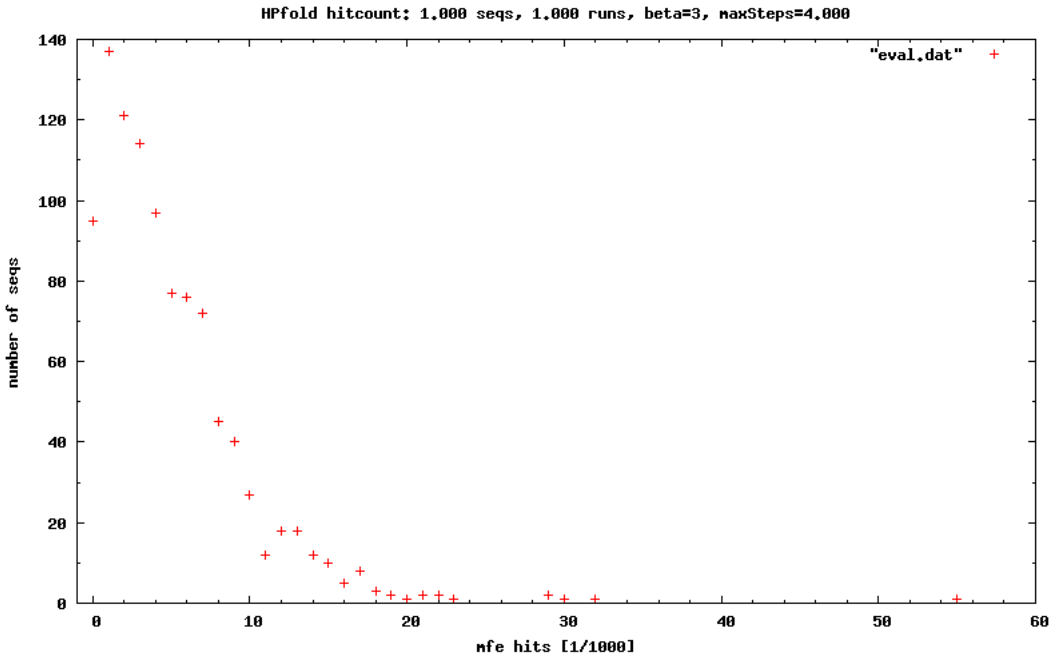

Abbildung 5: Trefferquote mit Anzahl der dazugehörigen Sequenzen. Für 97 Sequenzen wurde die optimale Struktur nie gefunden. Für genau einen Fund der optimalen Struktur steigt die Anzahl der Sequenzen auf 137, fällt für höhere Quoten aber rapide ab. Fünf der Sequenzen zeigen besonders gute Faltungseigenschaften mit jeweils 29, 30, 32 und 55 Treffern.

spricht demnach lediglich 0,5% der Ausgangsdaten. Um einen größeren Be-

reich als gute Falter kategorisieren zu können wäre es hilfreich, wenn sich das lokale Minimum bei 11 Treffern bei weiteren Untersuchungen als statistisch signifikant erweisen würde. Ausgehend von diesem ersten Datensatz könnten dann 0,99% der Sequenzen als gute Falter kategorisiert werden. Alternativ könnte willkürlich eine Grenze gezogen werden, um etwa die besten 50% als gute Falter, den Rest als schlechte Falter zu kategorisieren.

Literatur

[1] Lesh N, Mitzenmacher M and Whitesides SA. (2003) *A Complete and effective move set for simplified protein folding.* In: 7th An Int Conf on Res in Comp Mol Biol. RECOMB, Berlin.

[2] Wikipedia: Lattice Protein, geladen 3. März 2008

[3] Albrecht A, Steinhöfel K (2006) *Run-time Estimates for Protein Folding Simulation in the H-P Model.* In: 9th Int. Symp. on Artificial Intelligence and Mathematics, Fort Lauderdale, Florida.

[4] Wikipedia (de): Protein, geladen 3. März 2008

[5] Wikipedia (de): Protein-Bioinformatik, geladen 3. März 2008

[6] http://www.bioinf.uni-freiburg.de/Software/Libraries/