Seminararbeit

von
Michael Wachter
MNR 66361

im Rahmen des Seminars

CI-Methoden in der Bioinformatik

Prof. Eyke Hüllermeier

WS 2004/05

Das Thema der Ausarbeitung beruht auf dem Inhalt des folgenden Textes:

Steffen Schulze-Kremer
Application of Evolutionary Computation to Protein Folding with Specialized Operators
Evolutionary Computation in Bioinformatics, Kapitel 8, 2003.

– Dortmund, 31.01.2005 –

1 Einleitung

In dem Buch „Evolutionary Computation in Bioinformatics“ wird der Einsatz von CI-Methoden in der Bioinformatik anhand einiger ausgewählter Themen gezeigt. Dazu wird nach einer Einführung (Kapitel 1 und 2) jedes Thema in einem eigenen Kapitel behandelt.

Das hier betrachtete Kapitel ist Kapitel 8, ,,Application of Evolutionary Computation to Protein Folding with specialized Operators" von Steffen Schulze-Kremer.

Es wird ein genetischer Algorithmus vorgestellt, mit dessen Hilfe versucht wird, die tertiäre Struktur (Anordnung im Raum) von einem durch seine Arminosäuren-Sequenz gegebenen Protein zu berechnen. Dabei soll diese Struktur der natürlichen Struktur dieses Proteines möglichst nahe kommen.

2 Erster Ansatz eines GA

Für jeden evolutionären Algorithmus muss ein geeigneter Formalismus gefunden werden. In dem hier betrachteten Fall wurde ein ,,hybrider Ansatz" mit einem genetischen Algorithmus (GA) verwendet, der nicht wie normalerweise auf Bitstrings, sondern mit realwertigen Zahlen arbeitet.

Dieser Ansatz verfügt über eine Reihe Vor- und Nachteile. Vorteile sind zu einem die einfache Implementierbarkeit und die Möglichkeit, anwendungspezifische Variationsoperatoren zu verwenden. Allerdings sind auch 3 potentielle Nachteile zu sehen.

1. Die mathematische Grundlage von genetischen Algorithmen gilt streng genommen nur für Bitstring-Repräsentations.

2. Bitstring-Repräsentationen laufen in vielen Anwendungen schneller.

3. Ein zusätzlicher Codierungsschritt ist notwendig, um die Zahlen in Bitstrings umzuwandeln.

2.1 Repräsentation des Proteins

Das Protein kann auf verschiedene Arten repräsentiert werden. Gebräuchlich sind die kartesichen Koordinaten der einzelnen Atome oder Verdrehungswinkel zwischen den Atomen.

Die Benutzung von kartesichen Koordinaten erweist sich als schwierig. Hier kann es vorkommen, dass bei der Mutation von einigen Atompositionen diese nicht mehr den richtigen Abstand zu anderen Atomen haben oder sogar mit ihnen kollidieren. Daher muss für jedes Individuum geprüft werden, ob es eine sinnvolle Anordnung der einzelnen Atome darstellt, was erstens einen hohen Rechenaufwand erfordert und zweitens dazu führt, daß viele generierte Individuen verworfen werden müssen. Wegen dieses Nachteils wird sie hier nicht verwendet.

Die Repräsentation durch Torsionswinkel erweist sich dabei als sinnvoller.

Bei dieser Repräsentation werden alle Stellen im Molekül des Proteins betrachtet, an

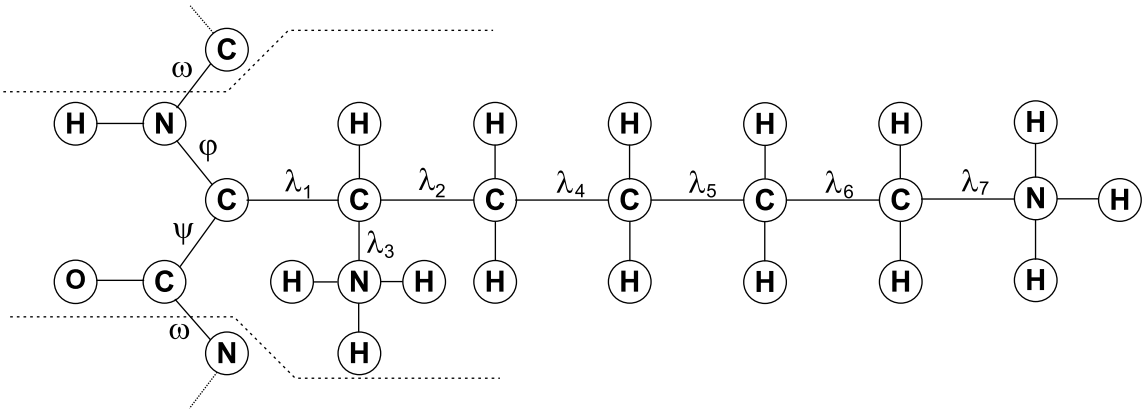

Abbildung 1: Ein Lysin-Residuum mit den Winkeln ϕ, ψ, und ω im Backbone und $\lambda_1..\lambda_7$ im Rest.

der die Anordnung der Atome durch Verdrehung verändert werden kann. Die Bindungsgeometrie, also die Länge der Bindungen und Bindungswinkel, werden dabei als konstant angenommen. Trotzdem erhält man genug Freiheitsgrade, um jede natürliche Anordnung von Proteinen mit nur kleiner Abweichung zu erhalten.

Dabei treten pro einzelner Arminosäure des Proteins die Winkel ϕ, ψ, ω, $\lambda_1..\lambda_7$ auf. Die Winkel ϕ, ψ, und ω des Backbones und λ_1 in der Bindung zum Rest gibt es dabei immer. Bei den in der Natur vorkommenden Arminosäuren können noch bis zu 6 weitere Winkel $\lambda_2..\lambda_7$ auftreten. (z.B. bei Lysin, Abbildung 1).

2.2 Fitness-Funktion

Das Programm CHARMm bietet eine Funktion, die für eine gegebene Proteinstruktur die Einzelenergien E_{bond}, E_{phi}, E_{tor}, E_{impr}, E_{vdW}, E_{el}, E_H, E_{cphi} berechnet. Mit der oben gemachten Vereinfachung der ausschliesslichen Nutzung der Torsionswinkel können einige dieser Energien als konstant angenommen werden, andere können vernachlässigt werden. Dadurch ergibt sich als Gesamtenergie

$$E_{ges} = E_{tor} + E_{vdW} + E_{el}$$

Versuche mit EAs zeigten, dass sich mit der Minimierung dieser Energie alleine Proteine nicht unbedingt in eine kompakte Form bringen lassen. Daher wird noch ein „ad-hoc pseudo-enthopic“ Term E_{pe} verwendet, der sich aus dem erwarteten und dem aktuellen Durchmesser des Proteins ergibt.

Der erwartete Durchmesser des Proteins wird als

$$\text{erwarteter Durchmesser} = 8 \cdot \sqrt[3]{\text{Länge}}$$

angenommen. Womit sich

$$E_{pe} = 4^{\text{aktueller Durchmesser}-\text{erwarteter Durchmesser}} \left[\frac{kcal}{mol}\right]$$

ergibt.

Das führt dazu, dass längliche Strukturen höhere Energiewerte erhalten als kugelförmige und somit kugelförmige bei der Minimierung bevorzugt werden.

Somit wurde die Gesamtenergiefunktion

$$E_{ges} = E_{tor} + E_{vdW} + E_{el} + E_{pe}$$

als Fitness-Funktion verwendet und minimiert.

2.3 Variationsoperatoren

Zur Veränderung der einzelnen Individuen des GAs wurden die Operatoren MUTATE, VARIATE und CROSSOVER entwickelt, die abhängig von einigen Parametern zur Laufzeit des Programmes ausgewählt werden.

2.3.1 MUTATE

Der MUTATE-Operator ersetzt einen Torsionswinkel duch einen zufällig aus den 10 am häufigsten vorkommenden Werten für diesen Torsionswinkel. Die Wahrscheinlichkeit, ob ein Torsionswinkel ersetzt wird, wird durch den MUTATE-Operator festgelegt. Welche die zehn häufigsten Torsionswinkelwerte sind, wurde durch eine statistische Analyse von 129 Proteinen herausgefunden.

2.3.2 VARIATE

Der VARIATE-Operator verändert einen Torsionswinkel um 1°, 5° oder 10°. Zuerst wird entschieden, ob der Winkel überhaupt, dann wird entschieden, um welchen Betrag und dann, in welche Richtung der Winkel verändert wird. Die Wahrscheinlichkeiten für diese Entscheidungen werden durch den VARIATE-Parameter und die Parameter für die drei verschiedenen Winkelparameter ausgedrückt.

2.3.3 CROSSOVER

Für den CROSSOVER-Operator werden 2 Individuen verwendet. Ein erster Parameter bestimmt die Wahrscheinlichkeit, ob diese beiden Individuen überhaupt verwendet werden. Zwei weitere Parameter bestimmen die Wahrscheinlichkeit, ob ein „two-point crossover“ und/oder ein „uniform crossover“ verwendet wird.

Beim „two-point crossover“ werden zufällig 2 Residuen eines Individuums ausgewählt. Dann wird der Teil des Proteins zwischen diesen beiden Residuen zwischen den beiden Individuen ausgewechselt.

Beim „uniform crossover“ wird für jedes Residuum zufällig ausgewählt, ob es vertauscht wird oder nicht.

2.3.4 Parametriesierung

Alle in den Variationsoperatoren vorkommenden Paramter ändern sich linear zwischen dem Beginn des Algorithmus und dem Abbruch nach 1000 Generationen. Dabei wer-

den die Wahrscheinlichkeiten für den MUTATE und der CROSSOVER Operator immer kleiner und die für den VARIATE-Operator immer größer. Beim VARIATE-Operator werden zu Beginn die großen Winkeländerungen, zum Ende hin die kleinen Winkeländerungen bevorzugt. Die Wahrscheinlichkeit des two-point crossover steigt bei zunehmender Zeit, die des uniform crossover fällt.

2.3.5 Selektion

Zur Selektion der Individuen für die nächste Generation wird „elitist-selection“ verwendet. Dabei überleben aus einer Population von 2n Individuen (n Eltern und n Nachkommen) die n Individuen mit dem besseren Fitness-Wert.

2.4 Ergebnisse

Der oben vorgestellte Algorithmus wurde mit dem Protein „Crambin“ getestet. Die natürliche Struktur von Crambin ist bekannt und es ist für einen Test recht gut geeignet, da es einerseits recht kurz ist (46 Residuen), andererseits aber einen starken amphiphilen Charakter hat, der es schwierig zu berechnen macht.

Bei diesem Test stellte sich heraus, das keines der Individuen der letzten Generation auch nur eine annähernde Ähnlichkeit mit der natürlichen Struktur hatte. Der GA hat total versagt.

Bei näherer Betrachtung der Energien stellt sich heraus, dass jede berechnete Struktur eine deutlich niedrigere Gesamtenergie als die natürliche auswies. Besonders die elektrostatische Energie E_{el} ist deutlich niedriger. Dies kann 3 verschiedene Ursachen haben:

- Elektrostatische Zusammenhänge tragen sehr viel mehr zur Stabilisierung bei.
- Crambin hat 6 teilweise geladene Residuen, die in dem Experiment nicht neutralisiert wurden.
- Das Minimieren der Gesamtenegie ist am einfachsten über das elektrostatische Potential möglich.

Da der GA eigentlich gut funktioniert und die Fitness-Funktion minimiert, ist diese nicht ausreichend, um eine der natürlichen Struktur nahekommende Struktur zu erzeugen. Daher wird eine bessere Fitness-Funktion benötigt, die später beschrieben wird.

Allerdings lässt sich die vorgestellte Fitness-Funktion gut zur Positionierung der Seitenketten nutzen. Dabei sind die Winkel ϕ, ψ, ω für einen gegebenen Backbone konstant und nur die Winkel λ_1 .. λ_7 werden optimiert. Die dabei mit dem GA ermittelten Strukturen stimmen sehr gut mit der natürlichen Strukture überein.

3 Verbesserung des GA

3.1 Zusätzliche Fitness-Kriterien

Da die Fitness-Funktion aus 2.2 keine brauchbaren Ergebnisse liefert, müssen zusätzliche Kriterien für die Fitness-Funktion gefunden werden.

Eine erste Fitness-Funktion verwendet nur die r.m.s - Abweichung der Struktur eines Individuums zu der natürlichen Struktur. Diese errechnet sich aus dem Abstand der Positionen von korrespondierenden Atomen beider Strukturen nach der Formel

$$\text{r.m.s} = \sqrt{\sum_{i}^{N} \left(|\bar{u}_i - \bar{v}_i|\right)^2}$$

Bei Werten zwischen 0Å und 3Å kann man von einer guten Übereinstimmung sprechen. Bei Werten zwischen 4Å und 6Å von einer geringen Übereinstimmung. Bei Werten über 6Å entsprechen sich nicht einmal mehr die Backbones der Strukturen.

Diese Fitness-Funktion kann dazu verwendet werden um zu prüfen, ob sich mit gegebenen Operatoren eine natürliche Struktur erzeugen lässt.

Eine zweite Fitness-Funktion wird als Vektor aus verschiedenen Fitness-Komponenten angenommen :

$$\text{Fitness} = \begin{pmatrix} r.m.s \\ E_{tor} \\ E_{vdW} \\ E_{el} \\ E_{pe} \\ polar \\ hydro \\ scatter \\ solvent \\ Crippen \\ clash \end{pmatrix}$$

Dabei sind E_{tor}, E_{vdW}, E_{el} und E_{pe} die Energien aus Kapitel 2.2. Die anderen Werte sind folgende:

- **polar**: Der Wert, der polare Residuen (Arg, Lys, Asn, Asp, Glu und Gln) an der Oberfläche bevorzugt. In Abhängigkeit von der Anzahl der Residuen N, der Anzahl der polaren Residuen und dem Abstand zum Schwerpunkt s ist dieser Wert

$$\text{polar} = \frac{-\sum_{i}^{N} |\bar{v}_i - s|}{k}$$

- **hydro**: Dieser Wert bevorzugt hydrophobe Residuen (z.B. Ala, Val, Il?, Leu, Phe, Pro, Trp) im inneren des Proteins.
- **scatter**: Mittlerer Abstand aller Residuen (C_α-Atom) zum Schwerpunkt des Proteins. Bevorzugt kompakte Strukturen.
- **solvent**: Oberfläche der Struktur, die „solvent-accessible“ ist.
- **Crippen**: Empirisch ermitteltes statistisches Potential. Summe über alle Paare von Atomen, die innerhalb einer gewissen Distanz miteinander interagieren.
- **clash**: Anzahl aller Atompaare, die näher als 3.8Å zusammen liegen. Dieser Wert kann zur Aproximation der Van Der Waals Energie E_{vdW} für kleine Distanzen verwendet werden.

Bei der späteren Verwendung wird der Fitness-Vektor mit einem Gewichtungsvektor multipliziert. Somit ist es möglich, den Einfluss der einzelnen Werte zu variieren und sogar einige Werte gar nicht zu berücksichtigen.

3.2 Spezialisierte Variationsoperatorn

Zur weiteren Verbesserung wurden noch einige Änderungen an den Variationsoperatoren durchgeführt :

3.2.1 Der LOCAL-TWIST Operator

Der LOCAL-TWIST Operator verändert ein drei Residuen langes Teilstück des Backbones so, dass Anfang und Ende an der gleichen Position bleiben.

Dafür werden Backbone-Winkel am ersten Residuum (ϕ_1,ψ_1) und am letzten Residuum (ϕ_2,ψ_2) des Teilstücks gesucht, die die Gleichung

$$u^+ T_\alpha R_{\phi_1} T_\beta R_{\psi_1+\pi} T_\alpha R_{\phi_2} T_\beta R_{\psi_2+\pi} T_\alpha e_1 - cos(\beta) = 0$$

erfüllen. Die Vektoren u, e_1, die Transformationsmatrizen T, die Rotationsmatrizen R und der Winkel β beschreiben dabei die Voraussetzungen für die ausschliesslich lokale Änderung.

3.2.2 Veränderungen am Mutate-Operator

Der Mutate-Operator wurde für die Backbone-Torsionswinkel ϕ und ψ verändert. Sie werden nun als Paare behandelt. Mittels eines clustering-Algorithmus wurden alle $\phi - \psi$-Paare aus 66 Proteinen in Cluster zu maximal 10 Paaren zusammengefasst. Der Mittelpunkt jedes Clusters (ein $\phi - \psi$ - Paar) kann nun zum Austausch der Winkel ϕ und ψ verwendet werden. Das führt dazu, dass bei der Anwendung des MUTATE-Operators die neuen Winkel ϕ und ψ besser zu in der natur vorkommenden Strukturen passen.

3.2.3 Verwendung der erwarteten sekundären Struktur

Wird eine gewisse sekundäre Struktur für einige Residuen erwartet (α-Helix oder β-Faltblatt), so kann der Wertebereich der Winkel ϕ und ψ eingeschränkt werden, was zu einer Verkleinerung des Suchraumes führt.

3.3 Ergebnisse

Benutzt man die gegebenen Variationsoperatoren MUTATE, VARIATE, CROSSOVER und LOCAL-TWIST mit der r.m.s - Fitness-Funktion so zeigt sich bei Crambin eine minimale r.m.s - Abweichung zu der natürlichen Struktur von nur 0,89Å. Dieser Wert liegt in der selben Größenordnung wie die Ungenauigkeiten bei der Vermessung der natürlichen Struktur. Somit eignet sich der GA-Ansatz mit diesen Variationsoperatoren gut zum Berechnen natürlicher Strukturen, falls eine passende Fitness-Funktion verwendet wird.

Allerdings kann die r.m.s - Abweichung nie 0 werden, da folgendes zu beachten ist:

1. Da die Bindungslängen und die Bindungsgeometrie bei dem Torsionswinkelansatz vernachlässigt wurden, kann nie genau die natürliche Struktur getroffen werden.
2. Die Operatoren MUTATE, VARIATE und CROSSOVER verursachen starke Änderungen der r.m.s - Abweichung, da eine Drehung an einem einzelnen Winkel im Backbone die ganze Struktur des Proteins ändern kann.
3. Nur der LOCAL-TWIST Operator kann kleine Verbesserungen der r.m.s - Abweichung erzeugen, da er die Struktur nur in einem begrenzten Stück verändert.

Versuche mit den Fitness-Komponenten polar, E_{pe}, E_{tor}, E_{el}, hydro, Crippen und solvent (ohne r.m.s) ergaben minimale r.m.s-Abweichungen von 6,27Å zu der natürlichen Struktur. Dabei wurde auch die Gewichtung der einzelnen Fitness-Komponenten verändert. Es zeigte sich, dass eine zu hohe Gewichtung von polar, E_{tor} und E_{el} zu Strukturen führt, die der natürlichen Struktur unähnlicher werden. Generell konnten durch die Veränderung der Gewichtung der Fitness-Komponenten keine geringere r.m.s-Abweichung als 6Å erreicht werden.

Weiteren Versuchen wurden mit den Fitness-Komponenten Crippe, clash, hydro und scatter gemacht. Zusätzlich wurden auch noch die Winkel ϕ und ψ unter Annahme der sekundären Struktur beschränkt. Der Winkel ω wurde auf 180° festgelegt. Durch diese Annahme konnte die r.m.s-Abweichung bei Crambin auf 4,36Å gesenkt werden.