Ein genetischer Algorithmus zur Proteinfaltung mit spezialisierten Operatoren

Michael Wachter
25. Februar 2005

Inhalt

Einleitung

Ein GA für die Proteinfaltung
- Repräsentation des Proteins
- Fittness-Funktion
- Variationsoperatoren
- Ergebnisse

Verbesserung des GA
- Fittness-Funktion
- Neue Variationsoperatoren
- Ergebnisse

Worum geht es bei der Proteinfaltung ?

- Primärstruktur des Proteins sagt nichts über die Anordnung im Raum.
- Funktion des Proteins ist abhängig von der Anordnung im Raum (Tertiärstruktur).
- Benutzung von evolutionären Algorithmen zur Bestimmung dieser Struktur.
- Ziel : Berechnen einer Struktur, die möglichst nahe an der natürlichen Struktur liegt.

Der Vortrag basiert auf Kapitel 8 „Application of Evolutionary Computation to Protein Folding“ von Steffen Schulze-Kremer aus dem Buch „Evolutionary Computation in Bioinformatics“

Ein GA für die Proteinfaltung

- Verwendung eines genetischen Algorithmus zur Bestimmung der tertiären Stuktur.
- Allerdings: keine Verwendung von Bitstrings, sondern Verwendung von realwertigen Zahlen.
- Vorteile :
  - Möglichkeit zur Verwendung anwendungsspezifischer Variationsoperatoren.
  - Einfach zu implementieren.
- Nachteile :
  - Mathematische Grundlage gilt streng genommen nur für Bitstrings.
  - Die Verwenung von Bitstrings ist schneller.
  - Die Zahlen müssen erst noch als Bitstrings codiert werden.

Repräsentation des Proteins durch Torsionswinkel

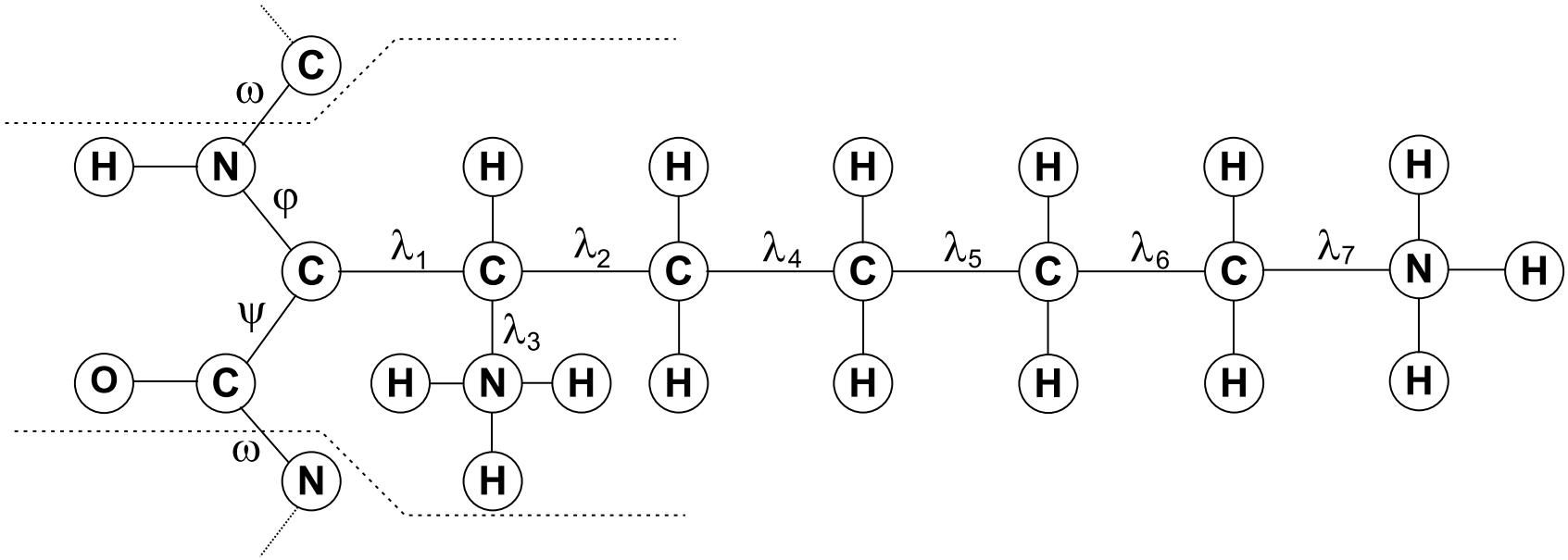

Abbildung: Lysin mit den Winkeln ϕ, ψ, und ω im Backbone und $\lambda_1..\lambda_7$ im Rest.

Fittness-Funktion

- Verwendung des Programms CHARMm zur Berechnung der Einzelenergien E_{bond}, E_{phi}, E_{tor}, E_{impr}, E_{vdW}, E_{el}, E_H, E_{cphi}
- Vernachlässigung der Energien E_{bond}, E_{phi}, E_{impr}, da sie beim Torsionswinkel-Ansatz konstant sind.
- Vernachlässigung der Energien E_H und E_{cphi} da das Protein im Vakuum betrachtet wird.
- Minimierung der Energie $E_{ges} = E_{tor} + E_{vdW} + E_{el}$.

Bevorzugung kompakter Strukturen

Problem: Diese Fittness-Funktion führt nicht zu kompakten Strukturen.

- Einführung einer zusätzlichen Pseudo-Energie, die kleiner ist, je kugelförmiger die Struktur ist.

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

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

- Verbesserte Energiefunktion:

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

Variationsoperatoren

- MUTATE
  Ein zufällig ausgewählter Torsionswinkel wird durch einen der 10 am häufigsten in der Natur vorkommenden ersetzt.
- VARIATE
  Ein Torsionswinkel wird mit einer gewissen Wahrscheinlichkeit um 1°, 5° oder 10° verändert. Durch Parameter wird angegeben, wie hoch die Wahrscheinlichkeiten für diese Winkel sind und in welche Richtung diese gedreht werden.

Variationsoperatoren

- CROSSOVER
  Auswahl von 2 verschiedenen Crossover-Varianten mit gewissen Wahrscheinlichkeiten
  - 2-Point Crossover - Die Torsionswinkel eines mehrere Residuen langen Teilstücks werden zwischen zwei Individuen vertauscht.
  - Uniform Crossover - Es werden die Torsionswinkel zweier korrespondierender Residuen mit einer gewissen Wahrscheinlichkeit vertauscht.

Parameter der Variationsoperatoren

- Alle Parameter ändern sich linear zwischen dem Beginn und dem Abbruch des Algorithmus nach 1000 Generationen.

Parameter	Wert Beginn	Wert Ende
MUTATE	80	20
VARIATE	20	70
VARIATE 10	60	0
VARIATE 5	30	20
VARIATE 1	10	80
CROSSOVER	70	10
CROSSOVER (uniform)	90	10
CROSSOVER (2-point)	10	90

Selektion

- Population von 10 Eltern und 10 Nachkommen.
- Die 10 besten Individuen überleben.

Ergebnisse

- Test des Algorithmus mit Crambin.
- Vorteil: Relativ kurz.
- Resultat: Keine auch nur annähernde Ähnlichkeit mit der natürlichen Struktur.
  Der Algorithmus hat hier komplett versagt.

Gründe für das Versagen

Der Wert für E_{el} ist deutlich niedriger als der der natürlichen Struktur.
Ursachen:

- Elektrostatische Zusammenhänge tragen sehr viel mehr zur Stabilisierung bei
- Crambin hat 6 teilweise geladene Residuen, die im Experiment nicht neutralisiert wurden
- Minimierung der Gesamtenergie ist am einfachsten über das elektrische Potential E_{el} möglich.

$\Rightarrow$Der Algorithmus hat die Fittness-Funktion gut optimiert, diese ist allerdings zur Lösung des Problems nicht geeignet.

Verbesserung der Fittness-Funktion

- Verwendung eines Fittness-Vektors aus verschiedenen Fittness-Komponenten.
- Möglichkeit der Gewichtung einzelner Fittness-Komponenten um bei Experimenten deren Einfluss ändern zu können.

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

Verschiedene Fittness-Komponenten

- E_{tor}, E_{vdW}, E_{el}, E_{pe} aus der ersten Fittness-Funktion
- r.m.s - Mittlere Abweichung der Struktur zu der natürlichen Struktur.
- polar - Wert, der Polare Residuen an der Oberfläche bevorzugt.
- hydro - Wert, der hydrophobe Residuen im Inneren bevorzugt.
- scatter - Mittlerer Abstand aller C_α zum Schwerpunkt der Struktur.
- Crippen - Anzahl von Atompaaren, die innerhalb einer gewissen Distanz miteinander interagieren.
- clash - Anzahl der Atompaare, die näher als 3,8Å zusammenliegen.

Verbesserung der Variationsoperatoren

- LOCAL-TWIST - Ein 3 Arminosäuren langes Teilstück des Proteins wird so verdreht, dass sich die Position der anderen Atome nicht ändert.
- MUTATE betrachtet nun die Winkel ϕ und ψ als Paar und ersetzt dieses Paar durch eins der 10 am häufigsten in der Natur vorkommenden Paare.
- Bei bekannter Sekundärstruktur kann man die Winkel ϕ und ψ weiter einschränken, was den Suchraum stark verkleinert.

Tauglichkeit der Variationsoperatoren

- Ausschliessliche Nutzung des r.m.s-Wertes für die Fittness-Funktion.
- Mittlere Abweichung nach 10000 Generationen: 0,89 Å
- Da die Bindungslängen und die Bindungsgeometrie vernachlässigt wurden, kann nie eine genaue Übereinstimmung mit der natürlichen Struktur erreicht werden.

Ergebnisse mit anderen Fittness-Komponenten

- Verwendung von E_{tor},E_{pe},E_{el},polar, hydro, Crippen und Solvent - minimale r.m.s - Abweichung von 6,27Å.
- Änderung der Gewichtung führte bis zu Abweichungen von 6Å.
- Verwendung von Crippe, clash, hydro und scatter mit konstanten Winkel ω und Beschränkung vom ϕ und ψ durch die sekundäre Struktur führt zu r.m.s - Abweichungen von 4,36Å

$\Rightarrow$ Diser Ansatz eines genetischen Algorithmus zur Proteinfaltung ist durchaus brauchbar.