# The Evolutionary Capacity of Protein Structures

Leonid Meyerguz, David Kempe, Jon Kleinberg, Ron Elber
RECOMB 2004

Alexander Hinneburg
Data Mining Seminar
Martin-Luther-Universität Halle-Wittenberg
SS 2006

# Einführung

- Beobachtung:

Proteinsequenzen

AJKLDHAOSHDNCUIHAORHNLJK
JKFJSKLDJFIkSDJFLKJSDLFJLKS
GRSKTDJFIHHOUIFHOIRNLKHSF
POIJOIJLFHSLDHFKLSHDIOHLHL

3D Proteinstruktur

- => hohe Redundanz bei der Abbildung Sequenz ➔ 3D-Struktur
- Protein-Struktur-Faltungsfamilien (ca. 500)

# Offene Fragen

- Redunanz zeigt wie Evolution den Raum der Proteinsequenzen durchmustert
- Fragen
  - Ist molekulare Evolution unvollständig?
  - Wurde erst ein kleiner Teil des Raumes der sinnvollen Sequenzen durchmustert?
- Anforderungen
  - Raum aller möglichen Sequenzen der Länge n über AS-Alphabet betrachten $AS = \{AS_1, \ldots, AS_{20}\}$
  - Approximative Algorithmen notwendig

# Techniken

- Analyse des Sequenzraumes $\mathcal{S}_n = AS^n$
  - Approximative kombinatorische Algorithmen
  - Realistische Energiefunktionen
  - Statistische Mechanik
- Fitness-Funktion $E_\sigma(X) : \mathcal{S}_n \rightarrow \mathbb{R}$
  - evaluiert zu einer Sequenz $X \in \mathcal{S}_n$ in einer festen 3D-Struktur $\sigma$ (Konformation) die Energie $E_\sigma(X)$
- Sequenz-Verteilungsfunktion
  $N(E) = |\{X : X \in \mathcal{S}_n,\ E_\sigma(X) \leq E\}|$
  - liefert die Anzahl von Sequenzen, welche in einer Konformation $\sigma$ eine Energie kleiner gleich *E* haben

# Kapazität und Temperatur

- Evolutionäre Kapazität
  - Sei eine natürliche Sequenz $E_\sigma$ in natürlicher Konformation $\sigma$ und $X_\sigma$ deren Energie
  - Dann ist $N(E_\sigma)$ die evolutinäre Kapazität, die ausdrückt wie weit die molekularen Evolution vom „engergetischem Optimum“ ist.
- Temperatur der Evolution
  - Trotz dessen, daß einige Sequenzen eine geringe Energie haben, gibt es viele mit deutlicher höherer Engergie, was ein Ausdruck der Entropie ist.
  - Mittels $N(E_\sigma)$ wird durch die statistische Mechanik ein analoges Maß zur Temperatur definiert, welche Energie und Entropie balanciert

# Evolutionäre Kapazität mit lokaler Fitness-Funktion

- Sei $X = x_1 x_2 \dots x_n$
- Lokale Fitness-Funktion $g(X) = \sum_{i=1}^{n} g_i(x_i)$
  - Beispiel THOM2 (Meller, J. and R. Elber (2001). Linear Optimization and a double statistical filter for protein threading protocols. Proteins, Structure, Function and Genetics 45: 241.)
- Evolutionäre Kapazität

$$N(E) = |\{X\colon X \in \mathcal{S}_n,\ g(X) \leq E\}|$$

# Verbindung zum Knappsack-Problem

- Vereinfachung
  - sei $\mathcal{S}'_n$ die Menge der Bitstrings der Länge n
  - N(E) und g(X) analog wie bisher
- Knapsack Problem
  - $n$ Objekte mit nicht-neg. Gewichten $a_1, \dots, a_n$
  - Zähle Bitvektoren $z = (z_1, \dots, z_n)$, s.d. $\sum_i a_i z_i \leq b$
  - Morris & Sinclair: Approx. Alg. mit Fehler $(1+\epsilon)$ und polynomiell in $n$ und $\epsilon^{-1}$

# Knapsack und Evolutionäre Kapazität

- Approx. Alg. für Sequenz-Verteilungsfkt. *N* über Bitstrings und lokaler Fitnessfkt. *g*
  - für $x_i \in \{0,1\}$ ist 0 die bessere und 1 die schlechtere Wahl, wenn $g_i(0) \leq g_i(1)$
  - Sei $E^*$ die Engergie der optimalen Sequenz $X^*$
    - für alle Positionen, die bessere Wahl
  - $z_i = 0$ entspricht besseren, $z_i = 1$ schlechteren Wahl
  - Knapsackgewichte seien $a_i = |g_i(0) - g_i(1)|$ und die Gesamtbeschränkung $b = E - E^*$
- Eine Lösung des Knapsackprob. entspricht gdw. einer Sequenz mit Energie $\leq E$, wenn für alle Pos. mit der besseren Wahl $z_i = 0$ und $g(X) \leq E$ .
- Algorithmus von Morris und Sinclair anwenden.

# Evolutionäre Kapazität für allg. Sequenzen

- Sei $x_i^*$ ein Symbol mit der min. Energie an Pos. i
- $N(E^*) = \prod_{i=1}^{n} k_i^*$ mit $k_i^*$ ist Anzahl der Symbole mit min. Engergie
- Angenommen
  - für die Energien $E^* = E_0 < E_1 < \ldots < E_m = E$ könnte $N(E_i)/N(E_{i-1})$ mit einem Fehler von $(1 + \epsilon/m)$ approximiert werden
  - dann wird $N(E)$ mit Fehler $(1 + \epsilon)$ durch folgendes Teleskopprodukt approximiert

$$N(E) = N(E_0) \cdot \frac{N(E_1)}{N(E_0)} \cdot \frac{N(E_2)}{N(E_1)} \cdots \frac{N(E_m)}{N(E_{m-1})}$$

# Approximation von $N(E_i)/N(E_{i-1})$

- Wähle Sequenzen gleichverteilt aus $\mathcal{S}_n^{(E_i)} = \{X \in \mathcal{S}_n : g(X) \leq E_i\}$
- Schätze $N(E_i)/N(E_{i-1})$ als reziproken Bruchteil der ausgewählten Sequenzen mit Energie $\leq E_{i-1}$
- $E_i$ und $E_{i-1}$ müssen nah genug sein, um hinreichend viele Sequenzen mit Energie $\leq E_{i-1}$ auszuwählen

# Gleichverteilte Auswahl aus $\mathcal{S}_n^{(E_i)}$

- Markov-Kette auf Sequenzen
  - Starte mit *X* und $E(X) = E^*$
  - Schritt: wähle Pos. i und Symbol $\alpha$
    - Sei $X' = x_1 x_2 \ldots x_{i-1} \alpha x_{i+1} \ldots x_n$
    - Falls $E(X') \leq E_i$ arbeite mit X' weiter
    - Sonst arbeite mit X weiter
- Theorem
  - Für jede Konstante $\delta > 0$ gibt es eine Anzahl von t Schritten, wobei t durch ein Polynom in n und $\log \delta^{-1}$ begrenzt ist, s.d. die Variationsdistanz zwischen der Gleich- und der Markov-Verteilung $\leq \delta$ ist für alle $t' > t$
- Alle t Schritt wird ein frisches Objekt fast gleichverteilt aus $\mathcal{S}_n^{(E_i)}$ erzeugt.

# Normalisierte Fitness-Funktion

- Befürchtung
  - lokale Fitness-Funktionen bevorzugen einfache Homo-Polymere
- Normalisierung
  - nutze $\hat{g}(X) = g(X) - g(X^{rev})$
  - Homo-Polymere und andere Sequenzen geringer Komplexität tendieren zu $\hat{g}(X) = 0$
  - Bei stark angepaßten Sequenzen ergibt sich oft $g(X) \ll g(X^{rev})$

# Nicht-Lokale Fitness-Funktionen

- Paarweise Fitness-Funktionen
  $g(X) = \sum_{i=1}^{n} g_i(x_i) + \sum_{i<j} g_{ij}(x_i, x_j)$
- Theorem
  - Für $k \geq 3$ gibt es paarweise Fitness-Fkt. für die ist es NP-hart zu entscheiden, ob $N(E) > 0$.
- Heuristik für paarweise Fitness-Fkt. TE13
  - Da Minimum-Energie Seq. nicht berechenbar
  - Telekop-Produkt von mittlerer Energie nach unten
  - Mittlere Energie durch direkte gleichverteilte Auswahl bestimmen

    $E = E_0 < E_1 < \ldots < E_m = \bar{E}$

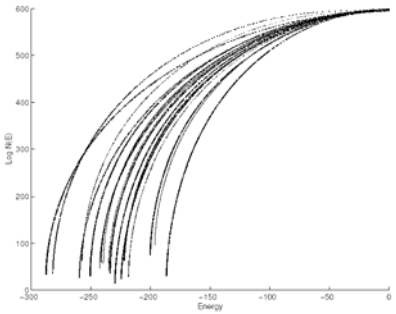
    $N(E) = \frac{N(E_0)}{N(E_1)} \cdot \frac{N(E_1)}{N(E_2)} \cdots \frac{N(E_{m-1})}{N(E_m)} \cdot N(E_m)$

# Experimente

- Daten
  - repräsentative Auswahl von 3409 Proteinenstrukturen aus der PDB
  - Proteine länger als 500 AS wurden weggelassen
- Für die approximativen Algorithmen wurden nicht die konservativen theoretischen Grenzen für t verwendet, sondern effizientere Standard-Heuristiken
- Die verschiedenen Energie-Funktionen THOM2, normalisierte THOM2, TE13 wurden verwendet

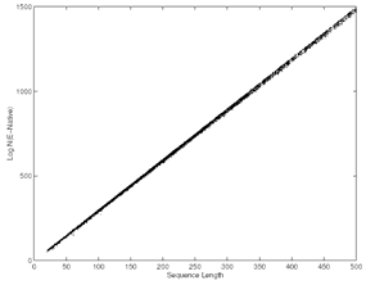
# Experimente

- log*N(E)* in Abhängigkeit von *E* für 19 verschiedene Proteine (n=199 o. n=200)
- Konvergenz bei $\log 20^n$
- Starker initialer Anstieg zeigt, daß dort der Sequenzraum viele Sequenzen mit niedriger Engergie enthält.




# Experimente

- log*N(E)* in Abhängigkeit von der Sequenzlänge für alle Proteine und Energiefkt.




- Trotz der unterschiedlichen Engergiefkt. weichen die Ergebnisse kaum von einander ab.

# Temperatur der Evolution

- Für kleine $\Delta E > 0$ , $\Omega(E) = \frac{N(E+\Delta E)-N(E)}{\Delta E}$ approximiert $\frac{dN(E)}{dE}$
- Sei $G(E)$ eine Fkt., welche die Überlebenswhr. einer Sequenz mit Engergie *E* zurückgibt.
- Die Whr. eine Sequenz zwischen $E$ und $E + \Delta E$ zu beobachten ist prop. zu $G(E)(N(E + \Delta E) - N(E)) \approx G(E)\Omega(E)\Delta E$
- Die zugehörige Whr.dichte ist $P(E) = G(E)\Omega(E)$
- Angenommen die nat. Seq. maximiert $P(E)$ (und so auch $\log P(E)$), dann gilt:

$$\frac{d[\log P]}{dE}|_{E_\sigma} = \frac{1}{\Omega(E)}\frac{d\Omega}{dE} + \frac{1}{G(E)}\frac{dG}{dE}|_{E_\sigma} = 0 \quad \Rightarrow$$

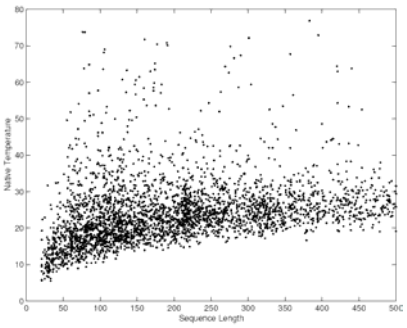
$$-\frac{1}{G(E)}\frac{dG}{dE}|_{E_\sigma} = \frac{1}{\Omega(E)}\frac{d\Omega}{dE}|_{E_\sigma}$$

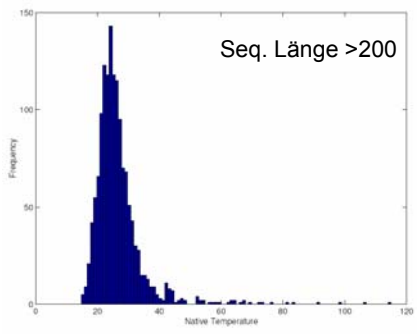
# Temperatur der Evolution (2)

- Gute Approximationen von N(E) ergeben Approximationen von $\Omega(E)$ und so kann wiederum $\frac{d\Omega}{dE}$ approximiert werden.
- Insgesamt kann so $-\frac{1}{G(E)}\frac{dG}{dE}$ für $E = E_\sigma$ approximiert werden was als $\beta_\sigma$ def. wird.
- Stat. Mechanik def. Temperatur als $\frac{1}{T} = \frac{dS}{dE}$ und mit $S = \log \Omega(E)$(Entropie) ist $\frac{1}{T} = \frac{1}{\Omega(E)}\frac{d\Omega}{dE}$
- D.h. $\beta_\sigma = \frac{1}{T_\sigma}$

# Ergebnisse

- Temperatur ist eine Funktion einer individuellen Proteinstruktur
- Es gibt keine Restriktion, warum die Temperatur für alle Strukturen gleich sein sollte
- Temperatur für alle Proteinstrukturen mit TE13




# Diskussion

- Die Temperatur für Proteinstrukuren mit mehr als 200 AS ist sehr konzentriert.
- => die allg. Selektionsfunktion G(E) verhält sich im Sequenzraum in der Nähe von nat. Sequenzen ähnlich für verschiedene Proteinstrukurfamilien.
- Mutationsprozess ist annähernd universal
- These wird gestützt durch zwei Modelle
  - Protein Cluster im Seq.Raum sind durch kleine Mutationsschritte verbunden
  - Ein einzelner Mutationsmechanismus kann ähnliche Verteilungen in verschiedenen Bereichen des Seq. Raumes erzeugen
- Fragen:
  - werden die Strukturen durch einen gleicharbeitenden Mutationsmechanismus erzeugt
  - oder durch direkte sehr langsame Migration zwischen den Proteinfamilien