Prof. Dr. Knut Reinert
Dr. Roland Krause
Matthias Winkelmann
Patrick Pett

Institut für Informatik
AG Algorithmische Bioinformatik

Algorithmische Bioinformatik
5. Übungsblatt WS 11/12

Abgabe bis Donnerstag, 24. November 2011, 15 Uhr

Aufgabe 1: Metropolis-Algorithmus
Wir wollen die Konvergenz und den Einfluss der Vorschlagsfunktion für das Erstellen einer Verteilung mittels des Metropolis-Algorithmus überprüfen.

Die Verteilungsfunktion ist hier mit $p(x) \propto 0.3 \cdot e^{(-0.2x^2)} + 0.7 \cdot e^{(-0.2(x-10)^2)}$ gegeben. Im Anwendungsfall ist sie natürlich nicht so einfach zu beschreiben, sonst könnte man sich die Darstellung mittels MCMC sparen.

Benutzen Sie die Vorschlagsfunktion (*proposal distribution*) $q(x^*|x^{(i)}) = \mathcal{N}(x^{(i)}, \sigma^2)$.

- Implementieren Sie den Metropolis-Algorithmus.
- Werten Sie Sie 5000 Iterationen für je $\sigma = 1, 10, 100, 1000$ aus.
- Stellen Sie die Ergebnisse in einem Streudiagramm (*scatter plot*) dar. Bewerten Sie Konvergenz nach 300, 1000 und 3000 Iterationen.
- Erzeugen Sie Histogramm der errechneten Verteilung. Ähneln Sie der gewünschten Verteilung? Wie lässt sich die Abweichung beschreiben?
- Wie unterscheiden sich Gibbs-Sampler, Metropolis- und Metropolis-Hastings-Algorithmus?

Aufgabe 2: PROJECTION-Algorithmus

Gegeben sind die vier Sequenzen $X_1 = ATGAGCA$, $X_2 = GTCGTGC$,$X_3 = GAGCACC$ und $X_4 = CTAGAGC$. Die Sequenzen enthalten ein Motiv der Länge 4 mit maximal einem Fehler.

- Wie hoch ist die Wahrscheinlichkeit, bei einer *bucket size* von 3 in einem Durchlauf des Algorithmus das gesuchte Motiv in allen Sequenzen zu identifizieren?
- Wie viele Durchläufe werden dann benötigt, um das gesuchte Motif mit 70% Wahrscheinlichkeit in allen Sequenzen zu identifizieren?
- Führen Sie einen Durchlauf des Algorithmus aus. Sie können dabei auf das Hashing der Bucket-Identifier und die Weiterverarbeitung mit dem EM-Algorithmus verzichten.

Aufgabe 3: Motif-Identifikation mittels phylogenetischen Signalen
Als Eingabsequenzen zur Identifikation von DNA-Motiven, wie sie in der Vorlesung vorgestellt wurden, werden typischerweise auf Promotor-Regionen innerhalb eines Organismus verwendet.

Regulatorische Sequenzen befinden sich meistens in intergenischen Bereichen, die weniger konserviert sind als protein-kodierende Bereiche (Gene). Wenn man also die Genome von verwandte Spezies zur Hand hat, kann man mittels vergleichender Genomik regulatorische Motive erkennen. Lesen Sie *Surveying Saccharomyces Genomes to Identify Functional Elements by Comparative DNA Sequence Analysis*[1] und fassen Sie die Arbeit zusammen.

a) Fassen Sie die Arbeit zusammen als ganzes in etwa 500 Worten zusammen.

b) Gehen Sie gesondert auf den Teil zu *Identification of Gene Regulatory Sequences* ein mit etwa 200 Worten ein. Sind die gefundenen Motive spezifisch oder in den Genomen der Hefe weitverbreitet?

Achten Sie auf sauber, wissenschaftliche Sprache. Vermeiden Sie Übersetzungsplagiate, also die wörtliche Übersetzung von zusammenfassenden Textteilen wie im Abstract oder der Einleitung.

[1] `http://dx.doi.org/10.1101/gr.182901`