technische universität dortmund

Bioinformatik, Fakultät für Informatik

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

Genominformatik, Medizinische Fakultät

Prof. Dr. Sven Rahmann

Sommersemester 2016

Algorithmische Bioinformatik
Übungsblatt 3

Ausgabe: 28. April 2016 · **Besprechung:** 12. Mai

Aufgabe 3.1 In dieser Aufgabe geht es um Proteine. Proteine sind Polymere aus Aminosäuren. Zur Vorbereitung empfiehlt sich eine Auffrischung der Grundlagen, zum Beispiel mit Hilfe des Vortrags "Protein Structure and Function" von Kristina Gremski (Teil 1: `http://www.youtube.com/watch?v=KH-LQSr7rHs`; Teil 2: `http://www.youtube.com/watch?v=MG8ziGyattk`).

Pfam (`http://pfam.sanger.ac.uk/`) ist eine Sammlung (Datenbank) von Proteindomänen und -familien, die durch Profil-HMMs repräsentiert werden. Besuche die Pfam-Webseite und beantworte folgende Fragen:

1. Was ist eine Proteindomäne; was ist eine Proteinfamilie; was ist ein Clan ?
2. Was ist der Unterschied zwischen Pfam-A und Pfam-B ?
3. Wir betrachten jetzt eine bestimmt Proteindomäne, die *Helix-Loop-Helix* (HLH) Domäne (manchmal auch "basic HLH (bHLH)" Motiv genannt), Pfam Accession number `PF00010`. In welcher Art von Proteinen findet man ein bHLH Motiv?
4. Betrachte ein multiples Sequenzalignment von mehreren Polypeptid-Sequenzen, die alle ein bHLH Motiv bilden (in Pfam über "Alignments" oder direkt: `http://pfam.sanger.ac.uk/family/PF00010/alignment/seed/html`). Das multpile Alignment beschreibt in jeder Spalte (die vorwiegend aus Aminosäurezeichen besteht) die Verteilung der Aminosäuren an einer bestimmten Position des Motivs. Ferner kann man sehen, dass manche Sequenzen manche Motiv-Positionen überspringen (Deletionen) und dass in manchen Sequenzen zusätzliche Aminosäuren eingefügt sind (Insertionen). Welche Schwierigkeiten ergeben sich, wenn man feststellen will, ob eine neue Sequenz in diese Familie gehört?
5. Um ein(e) Proteinfamilie (-domäne, -motiv) zu beschreiben, greift man gerne auf ein generatives Modell, insbesondere HMMs, zurück. Da Proteine Sequenzen sind, kommt vor allem eine lineare Modelltopologie in Frage, bei der es für jede Position des Motivs einen Match-Zustand und einen Deletions-Zustand, sowie zwischen zwei Positionen jeweils einen Insertions-Zustand gibt. Man spricht dann von *Profil-HMMs*; die Details sind in einem Artikel von Sean Eddy (1998) beschrieben: `http://www.ncbi.nlm.nih.gov/pubmed/9918945`. Lies den Artikel und beschreibe in eigenen Worten die Struktur eines Profil-HMMs für Proteinfamilien.
6. Eine praktische Art, ein Profil-HMM zu visualisieren, ist ein *HMM Logo*. HMM Logos stellen die Aminosäureverteilung an einer Position durch einen Stapel von Buchstaben dar. Insertions- und Deletionswahrscheinlichkeiten werden durch die Breite des Stapels und der Zwischenräume visualisiert; die Konserviertheit einer Position durch die Höhe. Betrachte das HMM Logo zu PF00010, lies ansatzweise den Artikel von Schuster-Böckler et. al (2004) über HMM Logos (`http://www.biomedcentral.com/1471-2105/5/7`), und beschreibe in eingenen Worten, wie genau die Höhe eines Stapels definiert ist.

Aufgabe 3.2 Der proteincodierende Bereich eines Gens in einem Prokaryoten (wie zum Beispiel des Bakteriums *E. coli*) beginnt mit einem Startcodon (oft ATG, steht auch für die Aminosäure Methionin) und endet mit einem Stoppcodon; die Codons dazwischen codieren die Aminosäuren des Proteins. Die Codons kommen nicht gleichverteilt vor, sondern mit einer bestimmten Häufigkeit, die auch von der Häufigkeit der jeweiligen Aminosäure abhängt (vgl. Aufgabe 1.2 von Blatt 1). Entwirf die Topologie eines *CDS finders* (coding sequence finders), also eines HMMs, das codierende Sequenzen in einem Genom erkennen kann.

Man braucht vermutlich je drei Zustände für ein Codon, irgend eine Struktur für die nichtkodierenden intergenischen Bereiche und sinnvolle Übergänge dazwischen. Kritisch betrachtet werden sollte die mit dem Modell erzeugte Längenverteilung einer codierenden Sequenz. (Dies ist keine Programmier-, sondern eine Designaufgabe.)

Aufgabe 3.3 Textmodelle und HMMs sind äquivalent.

Beweise die eine Richtung dieser Aussage, indem Du zu einem gegebenen HMM (Q, q_0, A, E, e) ein Textmodell (C, c_0, Σ, ϕ) angibst, so dass die gleichen Beobachtungs-Sequenzen mit gleichen Wahrscheinlichkeiten generiert werden.

Beweise die zweite Richtung dieser Aussage, indem Du zu einem gegebenen Textmodell (C, c_0, Σ, ϕ) ein HMM (Q, q_0, A, E, e) angibst, so dass die gleichen Beobachtungs-Sequenzen mit gleichen Wahrscheinlichkeiten generiert werden.

Es folgen noch zwei Aufgaben, die um besondere Beachtung bitten.

Aufgabe 3.4 Gehe am 05.05. nicht zur Vorlesung und Übung; es ist ein Feiertag! (Aus diesem Grund sind die Aufgaben auf diesem Blatt etwas langfristiger angelegt, und die Besprechung findet erst am 12. Mai statt.)

Aufgabe 3.5 Herr Decker bittet um möglichst viele Anmeldungen zur real-IT-y am 01.06.2016. Besuche die Webseite `http://www.real-IT-y.de` und entscheide, ob du dich anmelden möchtest.