Fachbereich Mathematik & Informatik
Freie Universität Berlin
AG Medical Bioinformatics
Tim Conrad

3. Praktikumsteil

Algorithmische Bioinformatik

WS 13/14

16. Dezember 2014

Abgabe: Sonntag, den 26.01.2014 bis 23:59 Uhr via Subversion.

Einführung

Ziel dieses Praktikums ist es, den Umgang mit Hidden Markov Modellen (HMM) zum Finden von Genen in DNA Sequenzen zu erlernen. Als Hilfsmittel empfehlen wir Ihnen

- BioJava (`http://biojava.org/wiki/Main_Page`)
- Jahmm (`http://code.google.com/p/jahmm/`)

Aufgabenstellung

1. Aufgabe — Daten sammeln

1. Laden Sie sich das *E. Coli* Genom im FASTA und GenBank Format herunter. `http://www.ncbi.nlm.nih.gov/nuccore/169887498`
2. Parsen Sie aus den GenBank Daten sämtliche Gene (Name, Position, etc.) auf dem positiven Strang in 3'-5' Richtung heraus.
3. Erstellen Sie mit Hilfe der Gen-Liste und der Gen-Sequenzen (aus der FASTA Datei) eine Liste mit allen kodieren und nicht-kodieren Bereichen des *E. Coli* Genoms.

2. Aufgabe — Erstellen, Trainieren und Benutzen eines HMM

1. Erzeugen Sie ein einfaches HMM (z.B. mit Hilfe der Jahmm API) zum Auffinden von Genen. Zwei Zustände (Gen, Nicht–Gen) sollten genügen.

2. Gegeben die Sequenz und die Genpositionen: implementieren Sie eine Methode, um Ihr HMM mit den Genen und der Sequenz zu trainieren. Benutzen Sie dazu die Häufigkeiten der Symbole und der Zustandsübergänge.

3. Implementieren Sie eine Methode, um ihr HMM mit dem Baum-Welch-Algorithmus zu trainieren.

4. Trainieren Sie Ihr HMM mit beiden Methoden und vergleichen Sie die geschätzten Parameter.

5. Trainieren und Validieren Sie das implementierte HMM mit einem der ihnen bekannten Kreuzvalidierungsverfahren (K-fold, Leave-One-Out, etc.). Benutzen Sie dabei zum Trainieren das häufigkeitsbasierte Verfahren.

3. Aufgabe — Steigern der Erkennungsrate durch ein komplexeres HMM

1. Erstellen Sie ein neues HMM, das mehr biologische Eigenheiten der Gene enthält (z.B. durch Modellierung des Start-Codons) und wiederholen Sie die Trainings- und Kreuzvalidierungschritte.

2. Vergleichen Sie die Ergebnisse der Kreuzvalidierung und die gefundenen Gene.

Auswertung

In dem abzugebenden Bericht sollten Sie auf verständliche Art und Weise die Implementierung Ihres Programms beschreiben. Beachten Sie, dass der Programmcode ebenfalls bewertet wird. Dieser sollte also ebenfalls gut kommentiert sein.

Beschreiben Sie das verwendete HMM und verdeutlichen Sie, warum Ihrer Meinung nach dieses HMM zum Auffinden von Genen geeignet ist. Beschreiben Sie Ihr Vorgehen beim Training des HMM. Erörtern Sie ihr Kreuzvalidierungsverfahren und die erhaltenen Ergebnisse. Begründen Sie die Erweiterung ihres HMMs aus der 3. Aufgabe.