

# Übungen zur Vorlesung Bioinformatik I

## Wintersemester 2003/2004

### Übungsblatt 6a

Abgabe: bis 28.11.03, 16.30 Uhr in INF 580

1. Sie möchten ein einfaches Programm schreiben, um codierende Sequenzen von nicht-codierenden zu unterscheiden. Gegeben seien ein Trainingsatz von ORF (open reading frames) Sequenzen, die jeweils ein durchgehendes Leseraster haben, und ein Programm, das die Häufigkeit der 64 Codons in beliebigen Sequenzen berechnet.

Geben Sie eine vernünftige Methode an, wie man codierende von nicht-codierenden Sequenzen unterscheiden kann. Die Methode soll statistisch fundiert sein. Wie erhalten Sie am einfachsten ein Modell für die Codon-Häufigkeit in nicht-codierenden Sequenzen?

(Hinweis: Erinnern Sie sich, wie man zufällige und nicht-zufällige Verteilungen vergleicht).

Wenn Sie die Berechnung tatsächlich ausprobieren möchten, benutzen Sie das PERL-Script `codon_usage2.pl`

([http://www.dkfz.de/ibios\\_old/lectures/bil\\_ws0304/codon\\_usage2.pl](http://www.dkfz.de/ibios_old/lectures/bil_ws0304/codon_usage2.pl)).

Um es zu benutzen, müssen Sie PERL installieren (<http://www.perl.com>). Das Programm rufen Sie aus einer MS-DOS Eingabeaufforderung auf (bzw. aus einer UNIX-shell, wenn Sie mit Linux arbeiten):

```
perl codon_usage2.pl <FASTA-Datei>
```

Eine FASTA-Datei mit Trainingssequenzen (77 ORFs aus dem Humanen Herpesvirus 1) finden Sie unter

[http://www.dkfz.de/ibios\\_old/lectures/bil\\_ws0304/hhv1.fasta](http://www.dkfz.de/ibios_old/lectures/bil_ws0304/hhv1.fasta).

(4 Punkte)