Bioinformatik I, WS00/01
Übungsblatt zu C/C++ - Programmierung

Besprechung am 9. November 2000

1 DNA-Generierung

Schreiben Sie ein C/C++-Programm, das zufällige DNA-Sequenzen erzeugt. Dabei soll die Wahrscheinlichkeit für jede Base A,C,G,T einzeln gewählt werden können, z.B.

$$p(A) = \frac{4}{15},\ p(C) = 0.2,\ p(G) = \frac{2}{15},\ p(T) = 0.4\,.$$

Überlegen Sie zunächst, wie man mit ungleichverteilten Wahrscheinlichkeiten auswählen kann, wenn man mit

`rand()/(RAND_MAX+1.0)`

nur gleichverteilte Zufallszahlen im Intervall $[0, 1[$ generieren kann.

2 DNA-Analyse

Schreiben Sie ein C/C++-Programm, das zu einer gegebenen DNA-Sequenz die folgenden Werte bestimmt (bzw. schätzt, siehe unten)

1. die Länge der DNA
2. die relativen Häufigkeiten f_x der Basen $x \in \{A, C, G, T\}$
3. die relativen Häufigkeiten $f_{x,y},\ x, y \in \{A, C, G, T\}$, der Paare aufeinanderfolgender Basen

Beachten Sie, das DNA-Sequenzen, die durch Sequenzierung gewonnen werden, in der Regel fehlerbehaftet sind. Das heisst, die Sequenzen enthalten Symbole N,Y,R,..., die für nicht exakt bestimmte Basen stehen. Finden Sie heraus (Literatur, Internet) was diese Symbole bedeuten (z.B. N steht für irgendeine Base) und überlegen Sie, wie mit diesen Basen statistisch korrekt umgegangen werden sollte. Versuchen Sie in ihrer Implementierung diese Symbole zu behandeln.

Zusatzaufgabe Um automatisch realistischen Input für dieses Programm zu generieren, kann man in einer Erweiterung des DNA-Generierungsprogramms auch diese Symbole für unvollständig spezifizierte Basen erzeugen (mit gegebenen Wahrscheinlichkeiten). Implementieren Sie diese Erweiterung und testen Sie damit das Analyse-Programm. Interpretieren Sie die Ergebnisse.