

## Algorithmen und Datenstrukturen in der Bioinformatik

### Siebente Praktikumsaufgabe WS 12/13

Abgabe Montag, 03.02., 15:00 Uhr

---

#### File I/O: Reversekomplement

---

Implementieren Sie ein Programm, welches Sequenzen im FASTA-Format<sup>1</sup> einliest und deren Reverskomplemente in gleicher Reihenfolge im FASTA-Format wieder ausgibt. Die Dateinamen der Ein- und Ausgabendatei werden über die Programmzeile angegeben (der dritte Parameter kann hier ignoriert werden). Sie dürfen nur die Hilfsmittel der Sprache selber und der STL benutzen um die Aufgabe zu lösen. Externe Bibliotheken wie SeqAn<sup>2</sup> oder Biopython<sup>3</sup> sollen nicht benutzt werden.

Beispiel:

```
ktrappe@peking:~$ ./aufgabe7 input.fa output.fa bin
```

**Zusatz:** Ist der dritte Programmzeilenparameter *bin* gesetzt, soll alternativ zur Standardausgabe die Datei in binär Format ausgegeben werden (**2 extra Punkte**).

Erweitern Sie die Funktionalität Ihres Programms um die FASTA-Dateien bitkomprimiert zu speichern. Dabei werden die einzelnen Buchstaben des DNA-Alphabets in kleineren Einheiten als 1 Byte abgespeichert. Da sie aber keine Bits direkt adressieren können, müssen Sie eine geeignete Funktion schreiben, die mit Hilfe von Bit-Operatoren Nukleotide bitweise in einem geeigneten Datentyp (zum Beispiel: char (1Byte), short (2 Byte), int (4 Byte) oder long long (8 Byte)) transformiert (**2 extra Punkte**).

---

<sup>1</sup><http://de.wikipedia.org/wiki/FASTA-Format>

<sup>2</sup><http://www.seqan.de/>

<sup>3</sup>[http://biopython.org/wiki/Main\\_Page](http://biopython.org/wiki/Main_Page)

Hinweise:

- Das Komplement einer Sequenz entsteht durch Austausch jedes Zeichens durch die Komplementärbase ( $A \leftrightarrow T$ ,  $C \leftrightarrow G$ ,  $N \leftrightarrow N$ ) (**2 Punkte**)
- Das Reverskomplement ist das Komplement von hinten nach vorn gelesen (**1 Punkt**)
- Die Kopfzeilen der Eingabedatei sollen genauso in der Ausgabedatei ausgegeben werden (**1 Punkt**)
- Kommentarzeilen sollen ignoriert, also nicht wieder ausgegeben werden (**1 Punkt**)
- In der Ausgabedatei sollen die jeweils letzten Sequenzzeilen höchstens 60 und alle anderen Sequenzzeilen genau 60 Zeichen enthalten (**1 Punkt**)
- Testen Sie ihr Programm ausgiebig auf selbst gewählten Eingaben
- Sie benötigen die ordValue-Funktion aus Aufgabe 5 für die Komprimierung (siehe: <https://svn.imp.fu-berlin.de/aldabi/WS12/material/aufgabe5/>)

Beachten Sie die Praktikumshinweise unter <https://www.mi.fu-berlin.de/w/ABI/AlDaBiWS12>.