Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2016/17
Martin Vingron· Annalisa Marsico · Alena van Bömmel · Edgar Steiger · Thimo Wellner

Blatt 1 · Ausgabe am 17.10.2016
Abgabe am 24.10.2016 vor Beginn der Vorlesung

Name: Matrikelnummer: Übungsgruppe:

Aufgabe 1 (*25 Punkte, Programmieren/Praxis*)**.** Laden Sie die Proteinsequenzen von *Saccharomyces cerevisiae* (Referenzsequenzen vom Stamm S288C), die Sie in der Saccharomyces Genome Database finden, herunter[1]. Die Proteinsequenzen sind unter Protein-Translationen der ORFs zu finden, verwenden Sie die Datei mit Translationen von allen systematisch genannten ORFs, außer „dubiosen" ORFs und sog. Pseudogenen. Die Datei hat ein FASTA-Format; lesen Sie die Beschreibung des FASTA-Formats[2], falls Sie mit dem Format nicht vertraut sind. Zur Vereinfachung der Datei schreiben Sie ein Programm, das die Kopfzeile jedes Eintrags durch die Gen-ID ersetzt. Ersetzen Sie also

```
>YAL001C TFC3 SGDID:S000000001, Chr I from 151006-147594,151166-151097,
Genome Release 64-2-1, reverse complement, Verified ORF, "Subunit of
RNA polymerase III  transcription initiation factor complex; part of the TauB
domain of TFIIIC that binds DNA at the BoxB promoter sites of tRNA and
similar genes; cooperates with Tfc6p in DNA binding; largest of six subunits
of the RNA polymerase III transcription initiation factor complex (TFIIIC)"
```

durch

```
>YAL001C
```

Die Proteinsequenzen sollen unverändert bleiben. Das Programm soll so gestaltet sein, dass es auf anderen Rechnern ohne Modifikationen laufen kann (Codes in Python bevorzugt). Die Ein- und Ausgabedateien sollen als Kommandozeilen-Parameter spezifiziert werden.

Aufgabe 2 (*20 Punkte, Praxis*)**.** Eine Forschungsgruppe hat bei einem Patienten einen krankheitsrelevanten Single Nucleotide Polymorphism (SNP) an der Position Chr7:148525904 (Assembly hg19) festgestellt.

A) Inspizieren Sie die Region um den SNP im UCSC Genome Browser[3]. Was ist bekannt über diese Region? Überlappt der SNP mit einem Gen (Exon/Intron)?

B) Im „common SNPs"-Track können Sie sehen, dass der SNP schon bekannt ist. Mit der zugehörigen ID können Sie in der dbSNP-Datenbank[4] mehr über den SNP erfahren. Welche Funktion (synonymous/missense) hat er? Welches Allel (A, C, G, T) ist das übliche und welches ist das mutierte an dieser Position? Mit welchem Krankheitsbild scheint der SNP zusammenzuhängen? Schauen Sie sich dazu die verlinkten Publikationen in Pubmed an.

[1] `http://www.yeastgenome.org/download-data/sequence`
[2] `https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp`
[3] `https://genome.ucsc.edu/`
[4] `http://www.ncbi.nlm.nih.gov/SNP/`

Aufgabe 3 (*15 Punkte, Programmieren*)**.** Generieren Sie eine zufällige DNA-Sequenz der Länge 1.000, wobei Sie annehmen können, dass die 4 Basen gleichverteilt sind. Nun mutieren Sie diese Sequenz über 10.000 Generationen jeweils an einer zufälligen Position. Das Programm soll ohne Modifikationen auf anderen Rechnern laufen (bevorzugt Codes in Python oder R).

Aufgabe 4 (*40 Punkte, Rechnen/Theorie, Phylogenie/Sequenzalignments*)**.** Betrachten Sie folgende Matrix, die während eines Alignments mittels dynamischer Programmierung entsteht.

A) Erläutern Sie das Prinzip der dynamischen Programmierung.

B) Welcher Algorithmus wurde verwendet, um die Matrix zu füllen?

C) Welche Scoring-Funktion (Matches, Mismatches, Gaps) liegt vor?

D) Füllen Sie die restlichen Felder der Matrix aus.

E) Geben Sie die optimalen lokalen Alignments der beiden Sequenzen an.

		T	T	C	G	G	A	A	C	G	T	T
	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	2	1	0	0	0	2	1	0	0
T	0	2	2	1	1	0	0	0	1	1	3	2
G	0	1	1	1	3	3	2	1	0	3	2	2
C	0	0	0	3	2	2	2	1	3	2	2	1
G	0	0	0	2	5	4	3	2	2	5	4	3
T	0	2	2	1	4	4	3	2	1	4	7	6
T	0	2	4	3	3	3	3	2	1	3	6	9
C	0	1	3	6	5	4	3	2	4	3	5	8
C	0	0	2	5	5	4	3	2	4	3	4	7
G	0	0	1	4	7	7	6	5	4	6		
G	0											
T	0											
		T	T	C	G	G	A	A	C	G	T	T