

Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2014/15

Martin Vingron · Juliane Perner · Annkatrin Bressin

Blatt 1 · Ausgabe am 13.10.2014

Abgabe am 20.10.2014 vor Beginn der Vorlesung

Name:

Matrikelnummer:

Übungsgruppe:

Aufgabe 1 (20 Punkte + 10 Bonuspunkte). 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. Berechnen Sie in jeder Generation die Hamming-Distanz zur Originalsequenz und stellen Sie deren Verlauf über die Generationen graphisch dar. Was sind ihre Beobachtungen?

Bonusfrage: Haben Sie eine Erklärung dafür?

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

1. Inspizieren Sie die Region um den SNP im UCSC Genome Browser¹. Was ist bekannt über diese Region? Überlappt der SNP mit einem Gen (Exon/Intron)?
2. Im 'common SNPs'-Track können Sie sehen, dass der SNP schon bekannt ist. Mit der zugehörigen ID können Sie in der dbSNP-Datenbank² mehr über den SNP erfahren. Welche Funktion (synonymous/missense) hat er? Was sind die bekannten Allele an dieser Position (A,C,G und/oder T)? Schauen Sie sich die Übersicht der verlinkten Publikationen in Pubmed an. Mit welchem Krankheitsbild scheint der SNP zusammenzuhängen?

Aufgabe 3 (20 Punkte). Sie sollen bekannte Informationen über die Funktion und die evolutionäre Geschichte des humanen Proteins 'EZH2' herausfinden.

1. Suchen Sie dazu den Eintrag für das humane EZH2-Protein in der Uniprot-Datenbank³. Schreiben Sie eine kurze Zusammenfassung seiner molekularen Funktion. Erwarten Sie anhand der bekannten Funktionen, dass das Protein spezifisch für den Menschen ist? Begründen Sie Ihre Antwort kurz.
2. Laden Sie nun die Fasta-Sequenz der kanonischen Isoform herunter und führen Sie mit dieser eine Protein-BLAST-Suche⁴ durch (Achtung: Setzen Sie die 'maximum target sequences' auf 20.000). Wie konserviert ist das Protein? Welche Proteindomänen wurden gefunden?

¹<https://genome.ucsc.edu/>

²<http://www.ncbi.nlm.nih.gov/SNP/>

³<http://www.uniprot.org/>

⁴www.ncbi.nlm.nih.gov/blast/

Aufgabe 4 (40 Punkte). Betrachten Sie die folgende Matrix, die während eines Alignments mittels dynamischer Programmierung entsteht.

1. Erläutern Sie das Prinzip der dynamischen Programmierung.
2. Welcher Algorithmus wurde verwendet, um die Matrix zu füllen?
3. Welche Scoring-Funktion (Matches, Mismatches, Gaps) liegt vor?
4. Füllen Sie die restlichen Felder der Matrix aus.
5. Geben Sie die optimalen lokalen Alignments der beiden Sequenzen an.

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