
Aufgabenblatt 13

HGP; BLAST/BLAT/Quasar; MSA

Abgabe: **08.02.2005** bis **17 Uhr** über Goya

Maximal: 20 Punkte

Namen nicht vergessen!

1. Nach dem Start des Human-Genome-Projects im Jahre 1990 kam 2001 die erste Meldung über eine fast vollständige Aufklärung des menschlichen Genoms (99%; 1997 waren erst drei Prozent entschlüsselt). Woran könnte es liegen, dass in den vergangenen vier Jahren trotzdem immer noch nicht das gesamte restliche Erbgut aufgeklärt werden konnte (derzeit $\approx 99.5\%$)? **3 P**
2. BLAT hält zur Beantwortung von Anfragen einen Index über das gesamte Genom im Speicher. Dieser Index besteht aus allen nicht zusammenhängenden k-mers. Wie lang sind diese k-mers bei DNA-BLAT und wie lang bei Protein-BLAT? **2 P**
3. Der Index von DNA-BLAT belegt etwa 1GB Platz, weswegen man ihn noch leicht im Hauptspeicher halten kann. Bei Protein-BLAT sind es immerhin 2GB. Versucht, diese Zahlen herzuleiten oder zu begründen! **2 P**
4. Stellt die jeweils wesentlichen Merkmale von BLAST, BLAT und Quasar kurz zusammen! (Voraussetzungen, Heuristiken, Performance, ..) **6 P**
5. Vervollständigt den Satz: "Bei erwarteter Sequenzähnlichkeit von $m=93\%$ findet Protein-BLAT in einer homologen Region T von 33 Aminosäuren praktisch immer einen perfekten Match der Länge mit einem Substring von Q"! **3 P**
6. Wir wollen für die drei Sequenzen AGGCGT, GCCGT und ACT ein *Multiple Sequence Alignment* finden. Erstellt dafür die Matrix/Matrizen (dynamische Programmierung in drei Dimensionen, Sum-of-Pairs; Match=0, Mismatch=1, Gap=1, lineare Gapkosten)! **4 P**