MATEMATISKA INSTITUTIONEN
STOCKHOLMS UNIVERSITET
Avd. Matematik

SJÄLVSTÄNDIGT ARBETE I MATEMATIK

Onsdagen den 10 september kl. 16.30 - 17.30 presenterar Josefine Röhss sitt arbete "Analysing k-mer distributions in a genome sequencing project" (15 högskolepoäng, grundnivå).

Handledare: Kristoffer Sahlin

Plats: Sal 32, hus 5, Kräftriket

Sammanfattning: The usage of next generation sequencing for de novo assembly of genomes is increasing rapidly. However, due to the short read length from the next generation sequencing protocols, the assembly process is complicated. Algorithms for genome assembly need to be developed further in order to obtain high quality results that meet the criteria for downstream analysis. One subject for improvement is choosing an optimal k-mer size, depending on certain other variables. This project examines how different genome and sequencing conditions such as the number of repeats in the genome, GC-content and coverage can affect the choice of k-mer size. We tested KmerGenie, a program designed to calculate the best value of k. We also developed programs that simulated a genome from a Markov chain and divided the genome into reads with which KmerGenie predicted the best value of k and developed our own program to divide the reads into k-mers and produce histograms that could be compared to the output from KmerGenie. We tested different values of coverage, GC-content and repeat content and all the outputs from KmerGenie were compared. The result from the tests show that KmerGenie is not always able to predict the best value of k. Depending on repeats in the genome and GC-content, the quality of the estimation can vary substantially.

Alla intresserade är välkomna!