Things to know before the final exam:

Perl:
- Basic data types and control structures (scalars, lists, for and while loops, if-else blocks)
- Basic string manipulation (functions used in the alignment program)

Global pairwise alignment:
- Definition of a sequence alignment
- Running time and space requirement of optimal pairwise sequence alignment by dynamic programming
- Score of an alignment with and without affine gaps

Dynamic programming algorithm for pairwise alignment
- Computation of scoring matrix V
- Traceback

Alignment accuracy
- Comparison of computed alignment to “true” one

Cross-validation
- Selection of gap penalty

Scoring matrix
- Computing substitution scoring matrix from alignments

BLAST:
- Basic BLAST algorithm
  - Find kmers by hashing query and target
  - Find maximal local segment
- Speedup in running time

Local alignment:
- Difference in recurrence and traceback between this and global alignment

Profiles:
- Score nucleotide against profile
- Profile vs profile

Multiple sequence alignment:
- Iterative heuristic
- Progressive alignment
HMMs
- Compute probability of alignment
- Viterbi
- Maximum likelihood estimation of probabilities
- Forward and backward probabilities
- Expected maximization (Baum Welch)

Expected accuracy
- Expected accuracy score
- Posterior probability calculation from a set of alignments

Short read mapping
- Masked seeds
- BFAST approach

Genome alignment
- General strategy
  - Find high scoring segments (HSPs)
  - Longest increasing subsequence of HSPs
  - Do constrained alignment in between HSPs

Applications:
- Short read alignment: limitations of existing programs
- Multiple protein sequence alignment: what is the state of the art approach?
- Genome alignment: limitations, accuracy
- Metagenomics

Metagenomics:
- Machine learning vs. alignment strategy
- Runtime and accuracy of both as given in CLARK paper

Practice problems:
1. Write a Perl script that computes the BLAST score between two sequences. The BLAST score is defined to be the number of common keys between the two sequences as given by a seed (which may be spaced).
2. Write a Perl script that computes the profile of a multiple sequence alignment
3. A simple algorithm for genome alignment
   a. Find anchors
   b. Then find longest increasing subsequence
4. Weighted Needleman-Wunsch
5. A simple strategy for metagenomics
6. Perl script for short read alignment
   a. First identify local region of genome to align fragment to
   b. Then perform Needleman-Wunsch or Smith-Waterman to obtain the full alignment