

CSE182 Study guide for Midterm

November 2, 2004

This is a sampling of questions from material presented in class. At the mid-term, you are allowed to bring an A4 size page (both sides) worth of class notes, but otherwise, it will be a closed book exam. Some of these questions are more open ended than what you would find in the mid-term. Bring a calculator to the class.

Algorithms:

1. Global/Local alignment
2. Space saving trick in alignment
3. Exact dictionary matching
4. Regular expression matching
5. Profile based scoring
6. MS-MS de novo interpretation
7. Viterbi algorithm for HMMs
8. Forward-Backward algorithms

Regular Expression Matching: Draw a finite automaton for the following partial PROSITE pattern.

C-x(1,3)-[LFY]-x(1)-[FYW]

Reg. Expression: Does the sequence $s = AACAAYYYNR$, have a substring that matches this PROSITE pattern from the previous question. Augment your automaton to allow substring queries. Now compute $N(j)$ for all positions j in s . Use the sets $N(j)$ to say why a substring of s contains the pattern.

Reg. Expression: Which tool in PROSITE allows you to search with PROSITE patterns. Which tool can you use to search with profiles?

Profiles: A profile can be converted into a Position Specific Scoring Matrix. Consider a Position specific scoring matrix PSSM P . For all amino-acids a , and all positions $1 \leq i \leq m$, $P[a, i]$ gives the score for a being in the i -th position. Assume a gap penalty g . Given a query sequence s , design an algorithm that checks if a substring of s aligns with the profile with a score greater than T .

BLAST: The default word size in BLAST is 11 for DNA, and 3 for protein sequences. Explain with an analysis, why the default word size is different.

Genomics: 40% of the human genome is composed of Repeat sequence. Specifically, there are a few short sequences that occur repeatedly in many copies. What are the implications of this for BLAST searches, in terms of performance, and P-value?

Protein Structure: Which of the two secondary structures is not stable by itself: alpha helix, or beta strand? How does it form a stable structure?

BLAST: A BLAST search of a query sequence of length 100 against a database of size 10^7 bp results in an alignment with bit-score = 278. What is the E-value of this alignment? What is the P-value?

BLAST: Consider a query of size 200bp and a database of size 10^6 base pairs. Explain, with some calculations, how much faster BLAST would be with a word size of 8. How does this change with a word size of 15. If BLAST is always faster with word-size 15, why do we want to use a smaller word-size?

BLAST: When you run BLAST, portions of the sequence are often masked out by X symbols. What does that mean?

tblastx and blastp both align two sequences at the protein level. How are they different?

Profiles: Consider the DNA position specific score matrix P shown below. Suppose we are interested in searching for sequences that score at least 10 in a gapless alignment against PSSM P . Give a list L of keywords of size 3, such that any sequence which scores at least 10 MUST contain at least one of the keywords in L . Note that this is not a frequency matrix (Profile). We have pre-computed the scores, so that the score for A in the 1st position is 5, and so on.

A	5	-1	-3	-2	5	3
C	-1	5	-1	3	2	-1
G	-2	2	4	-1	-4	-1
T	1	1	1	-3	-2	1

Seq. Alignment: Consider the sequences $s = GGAATCATTACCA$, and $t = AACCGATTCTGG$, a match score of 1, a mis-match score of -1 , and linear gap cost of -1 (No gap opening penalty), compute the best local and global alignments of s and t .

HMMs: Proline rich regions. Consider an HMM for detecting proline rich regions. It has two states, P (proline rich) and O (other), and a start state. No residue is emitted in the start state. Instead, we just jump to P or O with equal probability.

- Emission Probabilities: $e_P(P) = 0.5$, and $e_P(r) = \frac{1}{38}$ for all other residues. $e_A(r) = \frac{1}{20}$ for all residues r .
- Transition Probabilities: $A[s, P] = A[s, O] = 0.5$, $A[P, P] = A[O, O] = 0.7$, and $A[P, O] = A[O, P] = 0.3$.

Compute the Viterbi parse for $ACPACSPPPPPACP$.

Gene finding: What are splice-sites? Define transcription start? Define start of translation. Which amino-acid is present at the start of translation. Which codon signifies end of translation? (Note: A codon is the in-frame triplet that codes for an amino-acid).